

# Discovering Variability Patterns for Change Detection in Complex Phenotype Data

Corrado Loglisci<sup>1</sup>(✉), Bachir Balech<sup>2</sup>, and Donato Malerba<sup>1</sup>

<sup>1</sup> Dipartimento di Informatica, Università degli Studi di Bari “Aldo Moro”,  
Bari, Italy

{[corrado.loglisci](mailto:corrado.loglisci@uniba.it),[donato.malerba](mailto:donato.malerba@uniba.it)}@uniba.it

<sup>2</sup> Istituto di Biomembrane and Bioenergetica (IBBE), CNR, Bari, Italy  
[b.balech@ibbe.cnr.it](mailto:b.balech@ibbe.cnr.it)

**Abstract.** The phenotype is the result of a genotype expression in a given environment. Genetic and eventually protein mutations and/or environmental changes may affect the biological homeostasis leading to a pathological status of a normal phenotype. Studying the alterations of the phenotypes on a temporal basis becomes thus relevant and even determinant whether considering the biological re-assortment between the involved organisms and the cyclic nature of the pandemic outbreaks. In this paper, we present a computational solution that analyzes phenotype data in order to capture statistically evident changes emerged over time and track their repeatability. The proposed method adopts a model of analysis based on time-windows and relies on two kinds of patterns, *emerging* patterns and *variability* patterns. The first one models the changes in the phenotype detected between time-windows, while the second one models the changes in the phenotype replicated over time-windows. The application to Influenza A virus H1N1 subtype proves the usefulness of our *in silico* approach.

## 1 Introduction

Phenotype is the set of observable traits of an organism, such as biochemical, physiological and morphological properties, and it is the result of complex factors, especially the influence of environmental factors and random variation on gene expression. Health and disease conditions of an organism have often a direct effect on its phenotype. Therefore, studying changes in phenotypes enable researchers to identify the main events occurring at DNA and protein levels which are responsible of immediate phenotypic responses, such as increased viral virulence or site/s-specific drug resistance. Often, these events are caused by molecular evolution and can occur together, thus they can be interesting for a better comprehension of the disease, but their eventual interaction, in combination with their heterogeneous nature, makes even more complex the investigation in laboratory. The analysis of phenotype data becomes of great importance given the socio-economical impact of the related diseases. In this scenario, the use of technologies able to handle the huge amount of phenotype data, model

the complex and heterogeneous aspects of the phenotypes and elicit possible interactions can play a determinant role.

In this paper, we focus on the analysis of phenotypes and propose a data mining method in order to capture statistically evident changes emerged over time and track their repeatability. The proposed method adopts a model of analysis based on time-windows and uses a frequent pattern mining framework as mean for abstracting and summarizing the data. This enables us to search changes as differences between frequent patterns. Since frequency denotes regularity, patterns can provide empirical evidence about real changes. Frequent patterns are discovered from the phenotype data collected by time-windows and thus they reflect co-occurrences of gene expression data, protein sequences and epidemiological impact, which are frequent in specific intervals of time. The changes which can emerge in this setting regard differences between the frequent patterns of two time-windows. In particular, the changes we are interested in correspond to variations of the frequency of the patterns occurred from a time-window to the next time-window. Not all the changes are considered, but only those which are replicated over time. We extend the concept of *Emerging Patterns* in order to depict changes between two time-windows and introduce the notion of *Variability Patterns* in order to characterize changes repeated over time.

## 2 Basics and Definitions

Most of the methods reported in the literature represent phenotype data by using formalisms based on vectors or attribute-value sets, which model only global descriptive properties of the phenotypes (e.g., presence/absence of mutations). These solutions could be too limiting because they neglect the complex structure of the phenotypes and the inner relationships existing among the biological entities related to the phenotypes. To overcome this drawback, we use the (multi-) relational setting, which has been argued to be the most suitable formalism for representing complex data. In the relational setting, phenotypes and the related biological entities can play different roles in the analysis. We can distinguish them between target objects (*TOs*) and non-target objects (*NTOs*). The former are the main subjects of the analysis (i.e., phenotypes), while the latter are objects (i.e., DNA level sequences, protein mutations) relevant for the current problem and biologically associated with the former.

Let  $\{t_1 \dots t_n\}$  be a sequence of time-points. At each time-point  $t_i$ , a set of instances (*TOs*) is collected. A *time-window*  $\tau$  is a sequence of consecutive time-points  $\{t_i, \dots, t_j\}$  ( $t_1 \leq t_i, t_j \leq t_n$ ) which we denote as  $[t_i; t_{i+w}]$ . The width  $w$  of a time-window  $\tau = \{t_i, \dots, t_{i+w}\}$  is the number of time-points in  $\tau$ , i.e.  $w = j - i + 1$ . Two time-windows  $\tau$  and  $\tau'$  are *consecutive* if  $\tau = \{t_i, \dots, t_{i+w-1}\}$  and  $\tau' = \{t_{i+w} \dots t_{i+2w-1}\}$ . Two pairs of consecutive time-windows  $(\tau, \tau')$  and  $(\tau'', \tau''')$  are  *$\delta$ -separated* if  $(j + w) - (i + w) \leq \delta$  ( $\delta > 0, \delta \geq w$ ), with  $\tau = \{t_i, \dots, t_{i+w-1}\}$ ,  $\tau' = \{t_{i+w} \dots t_{i+2w-1}\}$ ,  $\tau'' = \{t_j \dots t_{j+w-1}\}$ , and  $\tau''' = \{t_{j+w} \dots t_{j+2w-1}\}$ . Two pairs of consecutive time-windows  $(\tau, \tau')$  and  $(\tau'', \tau''')$  are *chronologically ordered* if  $(j + w) > (i + w)$ . We assume that all the time-windows have the same

width  $w$  and we use the notation  $\tau_{h_k}$  to refer to a time-window and the notation  $(\tau_{h_1}, \tau_{h_2})$  to indicate a pair of consecutive time-windows.

Both *TOs* and *NTOs* can be represented in Datalog language as sets of ground atoms. A ground atom is an  $n$ -ary logic predicate symbol applied to  $n$  constants. We consider three categories of logic predicates: (1) *key predicate*, which identifies the *TOs*, (2) *property predicates*, which define the value taken by a property of a *TO* or of a *NTOs*, and (3) *structural predicates*, which relate *TOs* with their *NTOs* or relate the *NTOs* each other.

The following definitions are crucial for this work:

**Definition 1.** Relational pattern

A conjunction of atoms  $P = p_0(t_0^1), p_1(t_1^1, t_1^2), p_2(t_2^1, t_2^2), \dots, p_m(t_m^1, t_m^2)$ , is a relational pattern if  $p_0$  is the key predicate,  $p_i$ ,  $i = 1, \dots, m$  is either a structural predicate or a property predicate.

Terms  $t_i^j$  are either constants, which correspond to values of property predicates, or variables, which identify *TOs* or *NTOs*. Moreover, all variables are linked to the variable used in the key predicate [6].

A relational pattern  $P$  is characterized by a statistical parameter, namely the *support* (denoted as  $\text{sup}_{\tau_{h_k}}(P)$ ), which denotes the relative frequency of  $P$  in the time-window  $\tau_{h_k}$ . It is computed as the number of *TOs* of  $\tau_{h_k}$  in which  $P$  occurs divided the total number of *TOs* of  $\tau_{h_k}$ . When the support exceeds a minimum user-defined threshold  $\text{minSUP}$ ,  $P$  is said to be *frequent*.

**Definition 2.** Emerging pattern-EP

Let  $(\tau_{h_1}, \tau_{h_2})$  be a pair of consecutive time-windows;  $P$  be a frequent relational pattern in  $\tau_{h_1}$  and in  $\tau_{h_2}$ ;  $\text{sup}_{\tau_{h_1}}(P)$  and  $\text{sup}_{\tau_{h_2}}(P)$  be the support of the pattern  $P$  in  $\tau_{h_1}$  and in  $\tau_{h_2}$ .  $P$  is an emerging pattern in  $(\tau_{h_1}, \tau_{h_2})$  iff  $\frac{\text{sup}_{\tau_{h_1}}(P)}{\text{sup}_{\tau_{h_2}}(P)} \geq \text{minGR} \quad \vee \quad \frac{\text{sup}_{\tau_{h_2}}(P)}{\text{sup}_{\tau_{h_1}}(P)} \geq \text{minGR}$

where,  $\text{minGR}$  ( $> 1$ ) is a user-defined minimum threshold. The ratio  $\text{sup}_{\tau_{h_1}}(P)/\text{sup}_{\tau_{h_2}}(P)$  is denoted with  $\text{GR}_{\tau_{h_1}, \tau_{h_2}}(P)$  and it is called *growth-rate* of  $P$  from  $\tau_{h_1}$  to  $\tau_{h_2}$ . When  $\text{GR}_{\tau_{h_1}, \tau_{h_2}}(P)$  exceeds  $\text{minGR}$ , we have that the support of  $P$  decreases from  $\tau_{h_1}$  to  $\tau_{h_2}$  by a factor equal to the ratio  $\text{sup}_{\tau_{h_1}}(P)/\text{sup}_{\tau_{h_2}}(P)$ , while when  $\text{GR}_{\tau_{h_2}, \tau_{h_1}}(P)$  exceeds  $\text{minGR}$ , the support of  $P$  increases by a factor equal to  $\text{sup}_{\tau_{h_2}}(P)/\text{sup}_{\tau_{h_1}}(P)$ .

The concept of emerging patterns is not novel in the literature [2]. In its classical formulation, it refers to the values of support of the same pattern which has been discovered in two different classes of data, while here we extend it to represent the differences between the data collected in two intervals of time, and therefore, we refer to the values of support of the same pattern which has been discovered in two time-windows. In the following, we report an example of EP.

$P$ : *phenotype*( $P$ ), *clinical\_condition*( $P, C$ ), *dependent\_by*( $C, M$ ), *affects*( $M, N$ ).

with  $\tau_{h_1} = [1991; 1995]$ ,  $\tau_{h_2} = [1996; 2000]$ ,  $\text{sup}_{[1991; 1995]}(P) = 0.8$  and  $\text{sup}_{[1996; 2000]}(P) = 0.5$ . Here, the support of the pattern  $P$  decreases, whereby

of the growth-rate  $GR_{[1991;1995],[1996;2000]}(P)$  is 1.6 (0.8/0.5). By supposing that  $minGR=1.5$ , the pattern  $P$  is considered emerging in  $([1991;1995],[1996;2000])$ .

**Definition 3.** Variability pattern-VP

Let  $T : \langle (\tau_{i_1}, \tau_{i_2}), \dots, (\tau_{m_1}, \tau_{m_2}) \rangle$  be a set of chronologically ordered pairs of time-windows;  $P$  be an emerging pattern in all the pairs  $(\tau_{h_1}, \tau_{h_2})$  with  $h = i, \dots, m$ ;  $\langle GR_{\tau_{i_1}, \tau_{i_2}}, \dots, GR_{\tau_{m_1}, \tau_{m_2}} \rangle$  be the values of growth-rate of  $P$  in the pairs  $\langle (\tau_{i_1}, \tau_{i_2}), \dots, (\tau_{m_1}, \tau_{m_2}) \rangle$  respectively;  $\Theta_P : \mathbb{R} \rightarrow \Psi$  be a function which maps  $GR_{\tau_{h_1}, \tau_{h_2}}(P)$  into a discrete value  $\psi_{\tau_{h_1}, \tau_{h_2}} \in \Psi$  with  $h = i, \dots, m$ .  $P$  is a variability pattern iff:

1.  $|T| \geq minREP$
2.  $(\tau_{h_1}, \tau_{h_2})$  and  $(\tau_{k_1}, \tau_{k_2})$  are  $\delta$ -separated for all  $h = i, \dots, m-1$ ,  $k=h+1$  and there is no pair  $(\tau_{l_1}, \tau_{l_2})$ ,  $h < l$ , s.t.  $(\tau_{h_1}, \tau_{h_2})$  and  $(\tau_{l_1}, \tau_{l_2})$  are  $\delta$ -separated
3.  $\psi = \psi_{\tau_{i_1}, \tau_{i_2}} = \dots = \psi_{\tau_{m_1}, \tau_{m_2}}$

where  $minREP$  is a user-defined threshold.

A VP is a frequent pattern whose support increases (decreases) at least  $minREP$  times with an order of magnitude greater than  $minGR$ . Each change (increase/decrease) occurs within  $\delta$  time-points and it is characterized by the value  $\psi$ . Intuitively, a VP represents a variation of the frequency of the same pattern, which is manifested with a particular regularity. It is quite evident that discovering this kind of information is relevant for studies on epidemics and pandemics. An example of VP is reported here. Consider the following EPs

$P$ : *phenotype*( $P$ ), *clinical\_condition*( $P, C$ ), *dependent\_by*( $C, M$ ), *affects*( $M, N$ )  
emerging in  $([1991;1992],[1993;1994])$   
 $P$ : *phenotype*( $P$ ), *clinical\_condition*( $P, C$ ), *dependent\_by*( $C, M$ ), *affects*( $M, N$ )  
emerging in  $([1996;1997],[1998;1999])$   
 $P$ : *phenotype*( $P$ ), *clinical\_condition*( $P, C$ ), *dependent\_by*( $C, M$ ), *affects*( $M, N$ )  
emerging in  $([1999;2000],[2001;2002])$   
 $P$ : *phenotype*( $P$ ), *clinical\_condition*( $P, C$ ), *dependent\_by*( $C, M$ ), *affects*( $M, N$ )  
emerging in  $([2004;2005],[2006;2007])$

Here,  $\psi_{[1991;1992],[1993;1994]} = \psi_{[1996;1997],[1998;1999]} = \psi_{[2004;2005],[2006;2007]}$ ,  $\psi_{[1991;1992],[1993;1994]} \neq \psi_{[1999;2000],[2001;2002]}$ . By supposing  $minREP = 2$  and  $\delta = 6$ ,  $P$  is a variability pattern. Indeed,  $T : \langle ([1991;1992],[1993;1994]), ([1996;1997],[1998;1999]) \rangle$  meets the conditions (1) and (2) because  $|T| = 2$  and  $(1998-1993) < 6$ ; the discrete values of the growth-rate in  $([1991;1992],[1993;1994])$  and  $([1996;1997],[1998;1999])$  meet the condition (3). The pair of time-windows  $([1999;2000],[2001;2002])$  is not considered because  $\psi_{[1999;2000],[2001;2002]}$  does not meet the condition (3), while the pair of time-windows  $([2004;2005],[2006;2007])$  does not meet the condition (3) because  $(2006-1998) > 6$ .

### 3 The Algorithm

We propose an algorithm which discovers VPs incrementally as time goes by. It works on the succession  $\langle (\tau_{1_1}, \tau_{1_2}), \dots, (\tau_{h_1}, \tau_{h_2}), \dots \rangle$  of pairs of time-windows

obtained from  $\{t_1, \dots, t_n\}$ . Each time-window  $\tau_{u_v}$  (except that for the first and last one) is present in two pairs, that is, the pair  $(\tau_{h_1}, \tau_{h_2})$  where  $\tau_{u_v} = \tau_{h_2}$ , and the pair  $(\tau_{(h+1)_1}, \tau_{(h+1)_2})$  with  $\tau_{u_v} = \tau_{(h+1)_1}$ . This is done with the intent to capture the changes of support of the patterns from  $\tau_{h_1}$  to  $\tau_{u_v}$  and from  $\tau_{u_v}$  to  $\tau_{(h+1)_2}$ . For each pair of time-windows  $(\tau_{h_1}, \tau_{h_2})$ , the algorithm performs three steps: (1) Discovery of frequent patterns on the time-windows  $\tau_{h_1}$  and  $\tau_{h_2}$  separately; (2) Extraction of EPs by matching the frequent patterns discovered from  $\tau_{h_1}$  against the frequent patterns discovered from  $\tau_{h_2}$ . These EPs are stored in a pattern base, which is incrementally updated as the time-windows are processed; (3) Identification of VPs by testing the conditions of Definition 3 on the EPs stored in the base. Note that, when the algorithm processes the pair  $(\tau_{(h+1)_1}, \tau_{(h+1)_2})$ , it uses the frequent patterns of the time-window  $\tau_{h_2}$ , which had been discovered when the algorithm had processed the pair  $(\tau_{h_1}, \tau_{h_2})$ . This avoids of performing the step (1) twice on the same time-window. Details on these three steps are reported in the following.

### 3.1 Relational Frequent Pattern Discovery

Frequent patterns are mined from each time-window by using the method proposed in [5], which enables the discovery of patterns whose support exceeds *minSUP*. It explores level-by-level the lattice of the patterns, from the most general to the more specific ones, starting from the most general pattern (which contains only the key predicate). The lattice is organized according to a generality ordering based on the notion of  $\theta$ -subsumption [6]. Formally, given two relational patterns  $P1$  and  $P2$ ,  $P1$  ( $P2$ ) is more general (specific) than  $P2$  ( $P1$ ) under  $\theta$ -subsumption, denoted as  $P1 \geq_{\theta} P2$ , if and only if  $P2$   $\theta$ -subsumes  $P1$ , where  $P2$   $\theta$ -subsumes  $P1$  if and only if a substitution  $\theta$  exists such that  $P2 \subseteq P1$ . The method adopts a two-stepped procedure: (i) generation of candidate patterns with  $k$  atoms ( $k$ -th level) by using the frequent patterns with  $k-1$  atoms ( $k-1$ -th level); (ii) evaluation of the support of the patterns with  $k$  atoms.

The monotonicity property of the support value (i.e., a super-set of a non-frequent pattern cannot be frequent) is exploited to avoid the generation of non-frequent relational patterns. In fact, in accordance with the Definition 2, non-frequent patterns are not used for detecting changes and thus we can prune portions of the space containing non-frequent patterns. Thus, given two relational patterns  $P1$  and  $P2$  with  $P1 \geq_{\theta} P2$ , if  $P1$  is non-frequent in a time-window, then the support of  $P2$  is less than the threshold *minSUP* and it is non-frequent too in the same time-window. Therefore, we do not refine the patterns which are non-frequent.

### 3.2 Emerging Pattern Extraction

Once the frequent patterns have been discovered from the time-windows  $\tau_{h_1}$  and  $\tau_{h_2}$ , they are evaluated in order to check if the growth-rate exceeds the threshold *minGR*. Unfortunately, the monotonicity property does not hold for the growth-rate. In fact, given two frequent patterns  $P1$  and  $P2$  with  $P1 \geq_{\theta} P2$ , if  $P1$  is

not emerging, namely  $GR_{\tau_{h_1}, \tau_{h_2}}(P1) < minGR$  ( $GR_{\tau_{h_2}, \tau_{h_1}}(P1) < minGR$ ), then the pattern  $P2$  may or may not be an EP, namely its growth-rate could exceed the threshold  $minGR$ . However, we can equally optimize this step by avoiding the evaluation of the refinements of a pattern  $P$  discovered from the time-window  $\tau_{h_1}$  ( $\tau_{h_2}$ ) in the case  $P$  is non-frequent in the time-window  $\tau_{h_2}$  ( $\tau_{h_1}$ ). Note that this operation could exclude EPs with very high values of growth-rate (i.e., the strongest changes), but here we are interested in the changes exhibited by co-occurrences which are statistically evident in both intervals of time.

The EPs extracted on the pairs of time-windows are stored in the pattern base, which hence contains the frequent patterns that satisfy the constraint set by  $minGR$  on at least one pair of time-windows. Each EP is associated with two lists, named as *TWlist* and *GRlist*. *TWlist* is used to store the pairs of time-windows in which the growth-rate of the pattern exceeds  $minGR$ , while *GRlist* is used to store the corresponding values of growth-rate. To distinguish the changes due to the decrease of the support from those due to the increase, we store the values of growth-rate as negative when it decreases.

The base is maintained with two operations, namely insertion of the EPs and update of the lists *TWlist* and *GRlist* associated with the EPs. A pattern is inserted if it has not been recognized as emerging in the previous pairs of time-windows, while, if it has been previously inserted, we update the two lists.

### 3.3 Variability Pattern Identification

The step (3) works on the pattern base and filters out the EPs that do not meet the conditions of Definition 3. The function  $\Theta_P$  implements an equal-width discretization technique. It is applied to two sets of values obtained from the lists *GRlist* of all the stored EPs, the first set consists of all the positive values of growth-rate, the second one consists of all the negative values. Note that we have not infinite values of growth-rate because all the patterns considered are frequent, i.e., there are no values of support equal to zero. The ranges returned by the discretization technique correspond to the discrete values  $\psi_{\tau_h, \tau_{h+1}}$ . Thus, we have two sets of ranges  $\Psi^+$  and  $\Psi^-$ :  $\Psi^+$  refers to the discrete values obtained from the positive values of growth-rate, while  $\Psi^-$  refers to the discrete values obtained from the negative values. We replace the numeric values contained in the lists *GRlist* with the corresponding ranges in  $\Psi^+$  and  $\Psi^-$ . This allows us to obtain two separate sets of discrete values and capture the increases/decreases of the support of the patterns by representing them with a finite number of cases.

This new representation of the growth-rate could suggest to prune the EPs that are more general and conserve the EPs that are more specific when they have the same discrete values. But, this cannot be done because it is not guaranteed that there is equality between the discrete values over all the time-windows.

In this step, the algorithm performs two preliminary operations: (i) removal of the EPs where the lists *TWlist* and *GRlist* have length less than the threshold  $minREP$ ; (ii) sorting of the remaining lists *TWlist* by chronological order. The lists *GRlist* will be re-arranged accordingly.

The algorithm discovers VPs by working on the EP separately and it can identify more than one VP from a single EP. For each EP, it scans the *TWlist* once and incrementally builds the set  $T$  of each candidate VP. A candidate VP is characterized by one discrete value. During the scan, it evaluates the current pair of time-windows  $(\tau_h, \tau_{h+1})$  of *TWlist* and the relative discrete value  $\psi_{\tau_h, \tau_{h+1}}$  against with the latest pair of time-windows  $(\tau_k, \tau_{k+1})$  inserted in the set  $T$  of the candidate VP that has the same discrete value: if the pairs of time-windows are  $\delta$ -separated, then the pair  $(\tau_h, \tau_{h+1})$  is inserted in the set  $T$  of the candidate VP, otherwise it can be considered to start the construction of the set  $T$  of a new candidate VP having the same discrete value. Finally, the algorithm filters out the VPs with  $|T|$  less than the threshold *minREP*.

In order to clarify how the step (3) works, we report an explanatory example. Consider  $\Psi^+ = \{\psi', \psi''\}$ , *minREP*=3,  $\delta=13$  and the lists *TWlist* and *GRlist* built as follows:

$$\begin{aligned} TWlist : & \langle ([1970; 1972], [1973; 1975]), ([1976; 1978], [1979; 1981]), ([1982; 1984], [1985; 1987]), \\ & ([1988; 1990], [1991; 1993]), ([1994; 1996], [1997; 1999]), ([2010; 2012], [2013; 2015]) \rangle \\ GRlist : & \langle \psi', \psi', \psi'', \psi' \rangle \end{aligned}$$

By scanning the list *TWlist*, we can initialize the set  $T$  of a candidate VP' by using the pairs  $([1970; 1972], [1973; 1975])$  and  $([1976; 1978], [1979; 1981])$  since they are  $\delta$ -separated ( $1979-1973 < \delta$ ) and they have the same discrete value  $\psi'$ . The pair  $([1982; 1984], [1985; 1987])$  instead refers to a different discrete value ( $\psi''$ ) and therefore it cannot be inserted into  $T$  of VP'. We use it to initialize the set  $T$  of a new candidate VP'', which thus will include the time-windows referred to  $\psi''$ . Subsequently, the pair  $([1988; 1990], [1991; 1993])$  is inserted into  $T$  of VP' since its distance from the latest pair is less than  $\delta$  ( $1991-1979 < \delta$ ). Then,  $T$  of VP'' is updated with  $([1994; 1996], [1997; 1999])$  since  $1997-1985$  is less than  $\delta$ , while the pair  $([2010; 2012], [2013; 2015])$  cannot be inserted into  $T$  because the distance between 2013 and 1997 is greater than  $\delta$ . Thus, we use the pair  $([2010; 2012], [2013; 2015])$  to initialize the set  $T$  of a new candidate VP'''. The set  $T$  of VP' cannot be further updated, but, since its size exceeds *minREP*, we consider the candidate VP' as valid variability pattern. Finally, the candidate VP'' cannot be considered as valid since its size is less than *minREP*. The candidate VP''' is not even considered since  $|T_{\psi'}| < \text{minREP}$ .

## 4 A Case Study: Influenza A/H1N1 Virus

We performed an empirical evaluation on the phenotype data concerning the influenza A/H1N1 virus. The flu virus is a common cause of respiratory infection all over the world. The Influenza A virus can infect several species. This virus contains eight segments gene of negative single-stranded RNA (*PB2*, *PB1*, *PA*, *HA*, *NP*, *NA*, *M*, and *NS*) encoding for 11 proteins. The subtype of Influenza A virus is determined by the antigenicity (the capacity to induce an immune response) of the two surface glycoproteins, haemagglutinin (HA) and neuraminidase (NA) [3]. The subtypes circulating in the human populations

determining important clinical conditions are H1N1 and H3N2. They cause epidemics and pandemics by antigenic drift and antigenic shift, respectively.

The datasets we use comprise phenotype data describing isolate strains of viruses of three different species, i.e., human, avian and swine. These isolate strains have been registered from 1958 to September 2009, while the datasets have been generated as a view on Influence Research Database hosted at the NIAID BioHealthBase BRC<sup>1</sup> and contain 3221 isolate strains for human, 1119 isolate strains for swine, and 757 isolate strains for avian.

Experiments are performed to study the effect of the thresholds  $w$ ,  $\delta$ ,  $minGR$  and  $minREP$  on the discovered variability patterns and emerging patterns. The parameter  $minSUP$  is fixed to 0.1. In this case study, the time-points correspond to years, while the number of the ranges produced by the discretization function is fixed to 5. Statistics on the results are collected in Table 1.

In Table 1(a), we have the number of VPs and EPs when tuning  $w$  ( $\delta=20$ ,  $minGR = 2$ ,  $minREP = 3$ ). By increasing the width of the time-windows, the overall number of the time-windows decreases, which results in a shorter succession of pairs of time-windows where finding EPs and VPs. This explains the decrease of the number of EPs and VPs for swine and human. We have a different behavior for the phenotypes of avian, where the number of EPs and VPs increases. Indeed, the use of wider time-windows ( $w=10$  and  $15$  against  $w=5$  and  $7$ ) leads to collect greater sets phenotypes having likely higher changeability. In this case, it seems that phenotype change concerns longer periods. In Table 1(b), we have the results when tuning  $\delta$  ( $w=5$ ,  $minGR = 2$ ,  $minREP = 3$ ). As expected, higher values of  $\delta$  allow us to detect a more numerous set of VPs, which comprises both the replications of EPs which are closer and the replications of EPs that are distant. Whilst, when  $\delta$  is 10, we capture only the VPs that cover at most ten years. The threshold  $\delta$  does not affect the number of EPs since it operates after the extraction of the EPs. In Table 1(c), we have the results when tuning  $minGR$  ( $w=5$ ,  $\delta=20$ ,  $minREP = 3$ ). We observe that  $minGR$  has great effect on the number of VPs and on the number of EPs. Indeed, at high values of  $minGR$ , the algorithm is required to detect the strongest changes of support of the patterns, which leads to extract only the EPs with the higher values of growth-rate. This explains the decrease of the number of VPs. The threshold  $minREP$  has no effect on the EPs since it acts on the VPs only (Table 1(d),  $w=5$ ,  $\delta=20$ ,  $minGR = 2$ ). As expected, higher values of  $minREP$  lead to exclude the EPs that have a low number of replications. This means that the algorithm works on a smaller set of EPs, with the result to have a lower number of VPs. In particular, when  $minREP$  is 6, we have no VP that includes EPs repeated six times and distant at most 20 years.

In the following, we report an example of variability pattern discovered by the proposed algorithm from the human dataset with  $w=10$ ,  $\delta=20$ ,  $minGR = 2$ ,  $minREP=3$ : *phenotype(P), epidemiological\_condition(P,E), is\_a(E,enhanced\_Transmission\_to\_Human), dependent\_by(E,M1), is\_a(M1,mutation\_A199S), mutation\_of(M1,T), is\_a(T,protein\_PB2), dependent\_by(E,M2), is\_a(M2,mutation\_A661T)*,

<sup>1</sup> <http://www.fludb.org/brc/home.do?decorator=influenza>.



*mutation\_of(M2,T)*, *dependent\_by(E,M3)*, *is\_a(M3,mutation\_K702R)*, *mutation\_of(M3,T)*.

Here,  $T : \langle ([1958;1967], [1968;1977]), ([1968;1977], [1978;1987]), ([1988;1997], [1998;2009]) \rangle$ ,  $\psi = [2;3,5]$ . The pattern concerns the epidemiological condition ‘enhanced\_Transmission\_to\_Human’ with the mutations ‘A199S’ on the protein ‘PB2’ and the mutations ‘A661T’ and ‘K702R’. The frequency of this pattern increases three times by a factor included in the range  $[2;3,5]$ . This happens between the time-windows  $[1958;1967]$  and  $[1968;1977]$ ,  $[1968;1977]$  and  $[1978;1987]$ ,  $[1988;1997]$  and  $[1998;2009]$ . Virologists observed in swine a pattern similar to that illustrated above. Indeed, given that PB2 has an important role in viral replication, transcription and spread, the common pattern between human and swine can explain a possible reassortment event. This could be happened favored by this amino acid mutation allowing viral particles to be exchanged between the two host species.

**Table 1.** Total number of the variability patterns and emerging patterns discovered on the three species when tuning the width of the time-window  $w$  (a), the maximum admissible distance between consecutive pairs of time-windows  $\delta$  (b), minimum threshold of growth-rate  $minGR$  (c) and minimum threshold of repetitions  $minREP$  (d). In each cell, we have reported the statistics as number of VPs–number of the EPs.

|       | $w(\text{years})$ |        |       |      |       | $\delta(\text{years})$ |        |        |        |
|-------|-------------------|--------|-------|------|-------|------------------------|--------|--------|--------|
|       | 5                 | 7      | 10    | 15   |       | 10                     | 15     | 20     | 25     |
| swine | 11–126            | 11–126 | 8–63  | 2–18 | swine | 0–126                  | 6–126  | 11–126 | 15–126 |
| human | 20–176            | 20–176 | 10–69 | 2–44 | human | 14–176                 | 18–176 | 20–176 | 22–176 |
| avian | 1–4               | 1–4    | 2–10  | 2–10 | avian | 0–4                    | 0–4    | 1–4    | 1–4    |
| (a)   |                   |        |       |      | (b)   |                        |        |        |        |
|       | $minGR$           |        |       |      |       | $minREP$               |        |        |        |
|       | 2                 | 4      | 6     | 8    |       | 3                      | 4      | 5      | 6      |
| swine | 11–126            | 7–68   | 0–2   | 0–0  | swine | 11–126                 | 6–126  | 2–126  | 0–126  |
| human | 20–176            | 6–61   | 0–4   | 0–0  | human | 20–176                 | 7–176  | 3–176  | 0–176  |
| avian | 1–4               | 0–2    | 0–2   | 0–0  | avian | 0–6                    | 0–2    | 0–2    | 0–2    |
| (c)   |                   |        |       |      | (d)   |                        |        |        |        |

## 5 Related Works and Conclusions

Despite its relevance, the analysis of phenotypes variability is a problem that has attracted attention only recently. In biology, a vast research focused on the evolution of the phenotype in the genome, but very few attempts have been done for analyzing the evolution over short periods of time. Phenotype evolution has been addressed without linking it to genetic information. For instance, in [7], the authors consider the phenotype of patients as temporal clinical manifestations semantically annotated and propose a technique based on the constraint networks to automatically infer phenotype evolution patterns of generic patients. In this work, we mined the relevant proteomic information enclosing many epidemiological and pathogenic variable of Influenza virus to draw their history dynamics over defined time intervals. This comparative proteomics novel approach could emphasize the drivers sub-patterns (key traits) and link them

to phenotypes in an epidemiological study framework. It is noteworthy that our approach is complementary to the antigenic trees construction from genetic sequences since these trees use mainly the antigenic phenotypes information.

The use of frequent pattern mining for analyzing dynamic domains is of recent investigation in data mining research, while the repeatability of patterns over time seems novel. In [1], the authors designed a density-based clustering algorithm to detect novelties from complex data in streaming setting. Novelties are captured as patterns whose frequency significantly changes with respect to homogeneous clusters of frequencies computed on previous windows of the data stream. In [4], we investigated the problem of capturing structural changes within heterogeneous and dynamic networks with a new notion of emerging patterns revised to model changes local to the topology of the network.

In this paper we investigated the task of characterizing the temporal variability of complex phenotypes by defining a novel notion of patterns to track the repeatability of such variations. The experimental results highlight the strong influence of the input parameter of minimum growth-rate and minimum number of repetitions on the discovered variability patterns. Although the method has been applied to phenotype data, we plan to explore its viability also in other scenarios of life sciences.

**Acknowledgements.** The authors would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

## References

1. Ceci, M., Appice, A., Loglisci, C., Caruso, C., Fumarola, F., Malerba, D.: Novelty detection from evolving complex data streams with time windows. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 563–572. Springer, Heidelberg (2009)
2. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 43–52 (1999)
3. Furuse, Y., Suzuki, A., Kamigaki, T., Oshitani, H.: Evolution of the M gene of the influenza A virus in different host species: large-scale sequence analysis. *Virology* **6**(67), 67–79 (2009)
4. Loglisci, C., Ceci, M., Malerba, D.: Discovering evolution chains in dynamic networks. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) NFMCP 2012. LNCS, vol. 7765, pp. 185–199. Springer, Heidelberg (2013)
5. Loglisci, C., Ceci, M., Malerba, D.: Relational mining for discovering changes in evolving networks. *Neurocomputing* **150**(Part A), 265–288 (2015)
6. Plotkin, G.D.: A note on inductive generalization. *Mach. Intell.* **5**, 153–163 (1970)
7. Taboada, M., Alvarez, V., Martinez, D., Pilo, B., Robinson, P., Sobrido, M.: Summarizing phenotype evolution patterns from report cases. *J. Med. Syst.* **36**(Suppl 1), S25–S36 (2012)

<http://www.springer.com/978-3-319-25251-3>

Foundations of Intelligent Systems

22nd International Symposium, ISMIS 2015, Lyon,

France, October 21-23, 2015, Proceedings

Esposito, F.; Pivert, O.; Hacid, M.-S.; Rás, Z.W.; Ferilli, S.  
(Eds.)

2015, XXIII, 466 p. 78 illus. in color., Softcover

ISBN: 978-3-319-25251-3