Mining High Average-Utility Sequential Rules to Identify High-Utility Gene Expression Sequences in Longitudinal Human Studies

Alberto Segura-Delgado[a] (alberto.segura.delgado@gmail.com), Augusto Anguita-Ruiz[b] (augustoanguitaruiz@gmail.com), Rafael Alcalá[a] (alcala@decsai.ugr.es), Jesús Alcalá-Fdez[a] (jalcala@decsai.ugr.es)

[a] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain.
[b] Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, 18071 Granada, Spain.

Corresponding Author:
Jesús Alcalá-Fdez
Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain.
Tel: +34 958240429
Email: jalcala@decsai.ugr.es

Mining High Average-Utility Sequential Rules to Identify High-Utility Gene Expression Sequences in Longitudinal Human Studies

Alberto Segura-Delgado[a], Augusto Anguita-Ruiz[b], Rafael Alcalá[a], Jesús Alcalá-Fdez[a,*]

[a]*Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain.*
[b]*Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, 18071 Granada, Spain.*

Abstract

High-utility sequential pattern mining techniques have demonstrated good performance in identifying associations between mRNA levels in microarray experiments taking into account both the biological context of each gene and the temporal characteristics of the dataset. However, these patterns do not provide information about how likely it is that the events in the pattern occur in the order indicated, therefore causal relationships cannot be established between of them. This reduces their predictive ability, making difficult its direct applicability to the field of gene expression dynamic modelling. An alternative to sequential patterns which takes the confidence of the forecast into account is the discovery of sequential rules. Their natural and seamless relation to human behaviour makes them very suitable to understand complex models without missing the possibility of using the generated rules as a standalone prediction model. This contribution proposes an evolutionary algorithm optimizing multiple objectives for mining biologically relevant high average-utility sequential rules from longitudinal human gene expression data with a good compromise through average-utility and

*Corresponding author. Tel.: +34 958240429 E-mail: jalcala@decsai.ugr.es
Email addresses: alberto.segura.delgado@gmail.com (Alberto Segura-Delgado), augustoanguitaruiz@gmail.com (Augusto Anguita-Ruiz), alcala@decsai.ugr.es (Rafael Alcalá), jalcala@decsai.ugr.es (Jesús Alcalá-Fdez)

explainability. This proposal enhances the well-known NSGA-II to learn, by evolutionary optimization, the rules maximizing two objectives: Utility and Interestingness. Moreover, a restarting mechanism and an external population have been particularly designed and included in order to encourage diversity in the search process preserving all the rules found. The quality of our approaches has been analyzed using external biological resources, statistical analysis and comparing with other proposals from the literature.

Keywords: High Average-Utility Sequential Rules, Multi-Objective Evolutionary Algorithm, eXplainable Artificial Intelligence, Gene Expression Patterns, Obesity

---

1. Introduction

Biochemical reactions in humans are complex systems subjected to regulatory interactions between genes. In these genetics regulatory systems, there exists certain temporal lag in the way the genes relate with each other. In other words, some time may pass between the moment a gene protein is initially translated from its locus until it elicits a downregulation/upregulation of its target Liang & Kelemen (2017). Time delay gene regulation is particularly evident in longitudinal trials in humans. In such cases, identified modifications in the gene expression pattern of a tissue may represent the molecular consequence of external agents and provoke later modifications in the mRNA levels of other genes.

Genetic microarray platforms have enabled the creation of functional genomics data repositories, which provide a vast number of whole-genome expression profiles (Barrett et al., 2012). The increase in massive temporal microarray experiments opens up new possibilities, such as discovering interesting time-delayed gene-gene relationships. Association discovery is one of the most common Data Mining techniques (Han & Kamber, 2006) used to provide scientists with interesting and understandable knowledge that assists them in decision-making, in accordance with eXplainable Artificial Intelligence (XAI) (Barredo Arrieta et al., 2020; Fernandez et al., 2019). This concept of eXplainability is of special interest for molecular biologists, since a great part of its daily work consists of converting complex genetics regulatory systems into particular gene-gene interaction hypotheses susceptible of being tested and validated in wet-lab experiments.

Sequential pattern mining techniques have proven to be very effective in

extracting interesting frequent sequential associations between genes from time course omics data (Agrawal & Srikant, 1995; Pei et al., 2001; Yang et al., 2002; Fournier-Viger et al., 2017; Liu et al., 2017; Nasu et al., 2020). However, most of them assume that all genes have the same importance within the biological process. Clinical studies have shown that some genes tend to be more significant in that they cause a specific disease or are more effective in fighting it, therefore the frequency data alone may not be sufficient information to be able to identify interesting sequences in the trait of interest. Because of this, different studies consider a utility value for each gene, representing its importance within the biological process. The aim is to obtain gene expression sequential patterns with a high utility from a temporal microarray dataset (Dinh et al., 2018; Gan et al., 2021; Zhang et al., 2021; Zihayat et al., 2017). In addition, the length of these patterns has been taken into account in order to avoid itemsets that have an unreasonable estimated utility value, which obtains high average-utility patterns (Diop et al., 2020; Truong et al., 2019; Zhang et al., 2020). These patterns, however, do not provide any information on how likely it is that the elements in the pattern occur in the order indicated, and therefore they cannot establish causal relationships. This reduces their predictive ability, which presents some disadvantages when it is fully to the temporal gene networks modelling process.

Taking it into account, an alternative to sequential patterns which takes the confidence of the forecast into account is the discovery of sequential rules (Fournier-Viger et al., 2015). Rule based systems are transparent models understandable by itself without the need to make use of post-hoc explainability techniques (visualizations, text explanations, explanations by example, among others) to explain them, as it is the case with other models such as convolutional neural networks and self-organizing maps, among others. Their natural and seamless relation to human behaviour makes them very suitable to understand and explain complex models, without missing the possibility of using the generated rules as a standalone prediction model. Their ability to be simulated or thought by a human will be improved by reducing the coverage (amount) and the specificity (length) of rules generated (Barredo Arrieta et al., 2020).

Designing sequential rule mining algorithms which are able to address the utility of the data is a challenge at present for researchers. Because of that, different methods have been proposed for mining High-Utility Sequential Rules (HUSR) making use of different interesting measures (such as confidence and lift, among others) obtaining and sorting the rules based on

their inherent interest to the expert (Zida et al., 2015). These interesting measures allow us to avoid the limitations of sequential patterns. Although these patterns occur frequently, their confidence may be low and are rendered unhelpful when it comes to decision making or predictions. These rules are defined as the classical association rules ($X \rightarrow Y$, where $X \cap Y = \oslash$) (Agrawal et al., 1993) but, here, the antecedent event must happen prior to the consequent based on the sequential ordering structure of the temporal database. For instance, let us consider a simple rule $[gene_i \downarrow, gene_j \uparrow] \rightarrow$ (time delay) $[gene_z \downarrow]$ extracted from a temporal microarray dataset, which declares that the significant repression of gene i and the upregulation of gene j are followed by (or cause) the significant repression of gene z after a given time delay.

Genetic Algorithms or, in general, Evolutionary Algorithms (EAs) (Yu & Gen, 2012), are recognized as interesting methods which can be used to mine useful and understandable knowledge from temporal datasets (Chamazi & Motameni, 2019; Fernandez et al., 2019; Zhou & Hirasawa, 2019). These searching and optimization methods have been widely and effectively applied to address knowledge extraction tasks. They usually evolve a set of solutions (namely population) by making use of one evaluation criterion for learning the fittest problem solutions. Over the last years some studies have framed this task as a problem with multiple objectives in order to consider different performance criteria together, which may result in different equilibrium levels subject to the own problem and the observable knowledge type (Matthews et al., 2013). As such, Multiple Objective EAs (MOEAs) provide us with an interesting way to address this task (Deb, 2015), as it requires a set of equally valid solutions/rules to be generated with different trade-offs between the objectives/measures of interest in a single run, with each result tending to optimize one objective to a higher degree than another.

In this paper, we propose a new MOEA, called HAUS-rules, for mining sequential rules with a higher interest, easier to be interpreted and involving a higher average-utility from sequence datasets. To accomplish this, the well-known NSGA-II MOEA (Deb et al., 2002; Srivastava et al., 2021) is extended for evolutionary mining and antecedent condition selection of each sequential rule, maximizing two objectives: Utility and Interestingness. These objectives allows us to concentrate the search process on information of potential interest and usefulness to the user, resulting in a search space reduction which helps the genetic search technique to obtain more suitable rules. In addition, the length of the rules is taken into account in order to avoid the pitfall of the "long tail" (Diop et al., 2020) and to encourage rules that involve few

items/events, since a rule with a high number of antecedents and/or consequents might become difficult to interpret. This method is also improved by incorporating a restarting mechanism and an external population (EP) to encourage diversity in the search process and to preserve all the rules found without redundancies to avoid overlaps (reducing the number of rules and making it easier to interpret the rules obtained).

With the intention of evaluating the suitability of our proposal for the identification of relevant biological knowledge on human longitudinal data, we have used a temporal gene expression dataset (microarray), obtained from adipose tissue samples collected during a dietary weight-loss intervention in people with obesity (three temporal records available), considering several external biological resources (such as functional annotation and gene regulation datasets) to analyze the High Average-Utility Sequential Rules (HAUSRs) obtained. In addition, we experimentally test our proposal on 10 sequence datasets with utility values to assess the performance of our proposal. The quantity of sequences in these datasets is in the range of [730, 77512] and the quantity of events is in the range of [250, 13905]. Our proposal has been compared to another approach from the literature used to mine HUSRs (HUSRM (Zida et al., 2015)) and to a well-known algorithm used to mine sequential rules (CMRules (Fournier-Viger et al., 2012)). Some non-parametric statistical tests have been used to evaluate the performance of these algorithms over the 10 sequence datasets (Garcia et al., 2009).

This paper is organized as follows. Section 2 provides a brief description of the research problem and employed datasets, some concept definitions of HAUSRs and several quality metrics. The different components for the evolutionary learning of interesting and understandable HAUSRs are detailed in Section 3. Section 4 evaluates and discusses insights extracted from human temporal microarray data and the 10 sequence datasets. Some concluding remarks are summarized in Section 5.

2. Preliminaries

Firstly, the in vivo human temporal microarray data derived from the course of a dietary intervention is introduced in this section. Secondly, some concept definitions of HAUSRs and several quality metrics are briefly introduced. Table 1 shows a summary of the notations used.

Table 1: Summary of the notations used.

Notation	Description
s_i	Event sequence i
X, Y, I	Unordered itemsets
$\neg I$	Negation of the itemset I
$X \rightarrow Y$	Sequential rule: X e Y are unordered itemsets ($X \cap Y = \oslash$) and X must be located prior to Y based on the sequential ordering principle
$sup(I)$	Support of the itemset I
$conf(r)$	Confidence of the rule r
$minSup$	User-defined threshold's minimum support
$minConf$	User-defined threshold's minimum confidence
$minUti$	User-defined threshold's minimum utility
$util(I, s)$	Utility of an Itemset I in a sequence s
$aveUtility(r)$	Average-utility value of the rule r
CF	Certain Factor measure
$CF * aveUti(r)$	Product of the certain factor measure and the average utility of the rule r
$g_i \uparrow$	Upregulation of the gene i
$g_i -$	Nochange of the gene i
$g_i \downarrow$	Downregulation of the gene i
$nChrom$	Number of chromosomes in the population
$nEval$	Total number of evaluations
Mut_p	mutation probability
α	User-defined threshold's restarting
P_t	Initial population
CP_t	Candidate population

2.1. In vivo human temporal microarray experiment

In vivo temporal microarray data consists of longitudinal human gene expression data recorded during the development of 2 dietary programs (three temporal records available) for 57 obese individuals part of a long-term dietary trial (Vink et al., 2016) These subjects were randomly assigned to a group with a low-calorie diet (LCD) with 1250 kcal/day for 12 weeks, and a group with a very-low-calorie diet (VLCD) with 500 kcal/day for 5 weeks. In addition, both interventions were followed by a phase of weight stabilization for 4 weeks. Over the course of the dietary interventions, each subject provided an abdominal subcutaneous adipose tissue biopsy at 3 different time points (baseline, after weight loss and after the weight stabilization phase) (Vink et al., 2017). This dataset is publicly available in the online omics repository Gene Expression Omnibus (GEO) (ID:GSE77962) (Barrett et al., 2012).

Fluorescence intensity data from the interventions were transformed into

a numerical matrix of mRNA abundance values, in which the rows match to individuals and the columns match to tested gene probes. 22 individuals from the LCD and 24 from the VLCD presented valid gene expression data (refer to row values) and valid fluorescence measures for more than 33,000 probes (refer to column values). All collected temporal data points were combined into a matrix, with each individual presenting consecutive entries corresponding to the three microarray experiments performed (one DNA microarray per subject and time point). Thus, the number of rows definitely included in the matrix was 138.

A feature selection algorithm was applied to the knowledge-extraction process to reduce input probes, selecting only those that show a differential expression pattern between time points (within each experimental group). The differentially expressed probes are identified by testing gene expression changes between analyzed time points: weight loss period (from baseline to finish the weight loss phase), weight stabilization period (from the end of the weight loss phase to weight stabilization phase), and dietary trial period (from baseline to finish the weight loss phase). The selected probes present a Bonferroni-adjusted P-value < 0.05 and a signal log ratio ≥ 1 or ≤ -1 (*Log2(FoldChange)*) in a paired t-test with Bayesian correction (Sheskin, 2003). 431 probes were selected, matching 398 unique genes.

Then, data was normalized by means of the robust multichip average (RMA) algorithm (Irizarry et al., 2003)) in order to take advantage of the benefits of discretization in genetics: categories are more comprehensible for scientists and easier to use; and the different datasets are easier to homogenize in terms of interpretability (Gallo et al., 2016). A 431-probe discretized matrix was generated for each experimental group, in which each probe may belong to three categories: 0, 1, and 2, referring to the absence of change in gene expression between time points, downregulation and upregulation respectively.

Finally, one sequence dataset per diet group was created, in which each sequence corresponds to an individual. Each event in the event sequences provides us with discrete temporal information on the gene expression changes of a probe between analyzed time points. This information is represented by means of a code with 4 numbers; the first number refers to the direction of the change in gene expression levels (0 nochange, 1 downregulation, and 2 upregulation), while the rest of the numbers correspond to the identification of the i probe evaluated.

Table 2: A simple time course omics dataset with 5 subjects belonging to a long-term intervention.

Subjects	Sequences - Expression change events in 2 two time points
s_1	$\langle\{(g_1 \uparrow, 3)\ (g_2 \uparrow, 1)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2-, 0)\ (g_3 \downarrow, 3)\}\rangle$
s_2	$\langle\{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3-, 0)\}, \{(g_1 \downarrow, 1)\ (g_2-, 0)\ (g_3 \uparrow, 1)\}\rangle$
s_3	$\langle\{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3 \uparrow, 2)\}, \{(g_1 \downarrow, 1)\ (g_2 \uparrow, 1)\ (g_3-, 0)\}\rangle$
s_4	$\langle\{(g_1 \uparrow, 3)\ (g_2 \downarrow, 1)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2-, 0)\ (g_3 \downarrow, 3)\}\rangle$
s_5	$\langle\{(g_1 \uparrow, 3)\ (g_2-, 0)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3 \downarrow, 3)\}\rangle$

2.2. High Average-Utility Sequential Rules

In many application areas, the data collected consists of value/event sequences that are ordered linearly according to the time at which they occurred. This temporal information is essential to be able to extract useful knowledge from data and it is particulary important in omics science as it helps explain the regulatory mechanisms of biological processes. For this reason, sequential rules have been successfully used to solve different biological problems (Anguita-Ruiz et al., 2020; Kilgore et al., 2019; Nguyen et al., 2018). These rules are used to represent sequential dependencies between values/events (items) of a sequence dataset. These rules are in the form $X \rightarrow Y$, with X, Y, being unordered or ordered sets of items (itemsets) and by taking into account that $X \cap Y = \oslash$, and that X must be located prior to Y based on the sequential ordering principle (Fournier-Viger et al., 2015). The classic measures of support and confidence are normally used to analyze these rules. The itemset support I ($sup(I)$) is computed by division of the quantity of sequences containing I by the whole quantity of sequences in the sequence dataset. Thus, by taking into account a rule with form of $X \rightarrow Y$, the support and confidence measures can be defined as follows:

$$support(X \rightarrow Y) : sup(XY) \tag{1}$$

$$confidence(X \rightarrow Y) : \frac{sup(XY)}{sup(X)} \tag{2}$$

with $sup(XY)$ being the support of itemset XY (X must appear before Y in the set of sequences containing this itemset). For instance, let us consider a sequence dataset containing information from 5 individuals belonging to a long-term intervention ($s_1, s_2, \ldots, s_5$). Table 2 shows the expression change

events for each individual (with 3 possible status: nochange ($-$), downregulation ($\downarrow$), and upregulation ($\uparrow$)) of 3 genes (g_1, g_2, and g_3) in 2 two time points. Let us to suppose that we have the sequential rule $r : (g_1 \uparrow) \rightarrow (g_3 \downarrow)$, representing that the upregulation of gene 1 at the first time point is followed by (or causes) the significant repression of gene 3 at the second time point. The values of the rule r for the support and confidence measures are:

- $support(r) = \frac{3}{5} = 0.6$
- $confidence(r) = \frac{0.6}{0.6} = 1.0$

which shows that this sequential relationship between the genes 1 and 3 occurs in more than half of the subjects and with a confidence value of 100%. Many proposals for mining sequential rules attempt to discover rules whose support and confidence values are greater than the user-defined threshold's minimum support ($minSup$) and minimum confidence ($minConf$), but these measures do not satisfy some desirable properties (Berzal et al., 2002; Brin et al., 1997). The confidence is, for example, unable to identify negative dependence of the rule antecedent with the rule consequent since it is not taking into account the support of the rule consequent in the calculation.

Some authors have proposed other measures to help overcome the problems that occur when obtaining and sorting the rules based on their potential interest to the user (Geng & Hamilton, 2006). Table 3 presents some of the well-known measures for assessing the quality of the rules, such as Lift, Yule'sQ and CF. For instance, let us to consider the sequence dataset shown in Table 2 and the rule $r : (g_1 \uparrow) \rightarrow (g_3 \downarrow)$. The Lift, Yule'sQ and CF values for the rule r are calculated as follows:

- $Lift(r) = \frac{0.6}{0.6 \cdot 0.6} = 1.67$
- $Yules'Q(r) = \frac{(0.6 \cdot 0.4) - (0.0 \cdot 0.0)}{(0.6 \cdot 0.4) + (0.0 \cdot 0.0)} = 1.0$
- $CF(r) = \frac{1.0 - 0.6}{1.0 - 0.6} = 1.0$

which support the sequential relationship between the genes 1 and 3. Most of these algorithms usually extract sequential rules without paying particular attention to the utility of each item has in the problem. Nevertheless, the utility value allows us to focus the search process on sequential rules with items that are highly relevant to the expert in the problem of interest. The

Table 3: Summary of some measures to assess the quality of rules.

Definition	*Description*	*Domain*
$Lift(X \rightarrow Y)$ (Ramaswamy et al., 1998)		
$\frac{sup(XY)}{sup(X)sup(Y)}$	Relationship between confidence and awaited confidence: Measure < 1: Negative correlation betwwen X and Y Measure $= 1$: X and Y are independent Measure > 1: Positive correlation betwwen X and Y	$[0,\infty)$
$Yule's Q(X \rightarrow Y)$(Tan et al., 2002)		
$\frac{sup(XY)sup(\neg X\neg Y) - sup(X\neg Y)sup(\neg XY)}{sup(XY)sup(\neg X\neg Y) + sup(X\neg Y)sup(\neg XY)}$	Odds ratio: Measure < 0: Negative correlation betwwen X and Y Measure $= 0$: X and Y are independent Measure > 0: Positive correlation betwwen X and Y	[-1,1]
Certain Factor$(X \rightarrow Y)$(Shortliffe & Buchanan, 1975)		
if confidence$(X \rightarrow Y) >$ *sup(Y):* $\frac{confidence(X \rightarrow Y) - sup(Y)}{1 - sup(Y)}$ *if confidence*$(X \rightarrow Y) <$ *sup(Y):* $\frac{confidence(X \rightarrow Y) - sup(Y)}{sup(Y)}$ *Otherwise is* 0	Gain normalized: Measure < 0: Negative correlation betwwen X and Y Measure > 0: Positive correlation betwwen X and Y Measure $= 0$: X and Y are independent	[-1,1]

utility of an itemset I in a sequence s in which $(util(I, s))$ occurs is calculated as the sum of the utility in the sequence s of each item i of the itemset I (referred to as internal utility) by its importance or weight in the problem (referred to as external utility). Thus, by considering a rule with form of $X \rightarrow Y$, the utility can be defined as follows:

$$utility(X \rightarrow Y) = \sum_{s \in D} util(XY, s) \tag{3}$$

where D is a sequence dataset. A number of proposals attempt to generate a set of HUSRs whose utility value is greater than the user-defined threshold's minimum utility ($minUti$) or the set with the extracted top-k HUSRs. However, the itemset length (number of items that it contains) may cause a noticeable effect on their utility value. Itemsets that present a high utility may give rise to unuseful rules since they could have been extended with items with high support and low utility. For this reason, the utility is usually divided by the itemset length to obtain high average-utility rules ($aveUtility(r)$) (Hong et al., 2011; Wu et al., 2018). For instance, let us to consider sequence dataset above with 5 subjects belonging to a long-term intervention. Table 2 shows the speed of change that each gene experiences in each sequence at each time point, and Table 4 presents for each gene its external utility, which considers the number of publications in which there is

Table 4: External utility of each gene in the time course omics dataset for the long-term intervention.

Gene	g_1	g_2	g_3
Utility	8	2	9

a direct relationship between the gene and obesity. The average-utility value of the rule $r : (g_1 \uparrow) \rightarrow (g_3 \downarrow)$ is calculated as follows:

- $aveUtility(r) = \frac{3 \cdot ((3 \cdot 8) + (3 \cdot 9))}{2} = 76.5$

In addition, the interestingness measures may be combined with utility functions to evaluate their potential interest to the user, considering the information provided by both types of measures (Diop et al., 2020). For example, obtained rules may be selected and ranked by the product of the Certain Factor (CF) and the average utility ($CF * aveUti$) . The HAUSRs obtained in the experimentation presented in this paper will be evaluated using the measures described in this subsection. Thus, the $CF * aveUti$ value of the rule $r : (g_1 \uparrow) \rightarrow (g_3 \downarrow)$ is calculated as follows:

- $CF * aveUti(r) = 1.0 \cdot 76.5 = 76.5$

3. New MOEA for Mining HAUSRs

We extent the well-known MOEA NSGA-II (Deb et al., 2002) for mining a set of sequential rules with a higher interest, easier to be interpreted and involving a higher average-utility, including two new components (EP and restarting mechanics) and maximizing two objectives: Utility and Interestingness. In this section, we will explain all the evolutionary learning components of our proposal in detail. Figure1 shows a workflow of the learning process to obtain HAUSRs.

3.1. Chromosome coding and initial population

A succession of genes represents a chromosome, which depicts those events involved in the sequential rule. Each gene has two parts:

- The first part (ev) depicts the involved event.

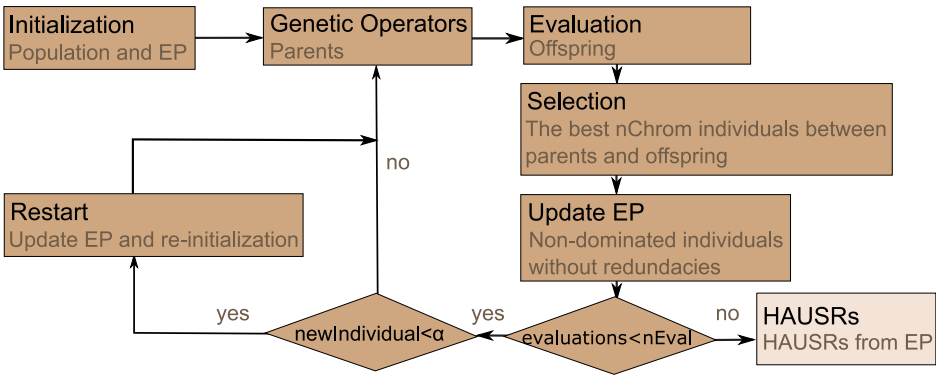

Figure 1: Workflow of HAUS-rules.

- The second part (ac) indicates if the event is part of the antecedent or of the consequent of the rule. When this part is '0', this event is part of the antecedent, and when this part is '1' is part of the consequent, taking the order of events in the rules into account.

Thus, a chromosome C is represented as follows, where n is the number of events involved in the rule:

$$Gene_i = (ev_i, ac_i), \;\; i = 1, \ldots, n \;\; ,$$
$$C = Gene_1 Gene_2 \ldots Gene_n$$

The population initially consists of a set of sequential rules with only one item/event in the consequent, so as to get simple rules that are easy to understand by the expert, even though this scheme allows many genes to be considered as a consequent part. In order to generate the initial gene pool (or individual population), we randomly choose an unmarked sequence s from dataset and then we select the events from it based on the application of a tournament selection attending to the utility of each event. It will be partially used for rule antecedent and rule consequent. To prevent generating rules with many events involved, the maximum event number that may be selected from the s is limited to $maxEvent$ (which is determined by the expert, usually at 5). Then, the sequences of the dataset that contain the generated sequential rule are marked. This mechanism is iteratively applied to unmarked sequences from dataset until the initial population is completely generated. In addition, when no unmarked secuences remain and the generated population is still not complete, the marks for all the sequences are removed and this mechanism is reapplied until the population is completely

generated. This mechanism allows information to be obtained from the whole dataset.

3.2. Objectives

Two objectives are maximized to obtain the sequential rules: Utility and Interestingness. The objective utility allows us to focus the search process on the information that is most useful to the user. With these aims in mind, we have used the utility measure *average-utility*, which takes the length of the rule into account when evaluating its utility. This measure allows us to avoid the pitfall of the "long tail" and encourages rules that involve few items/events (Diop et al., 2020). Notice that, the greater the quantity of events considered in order to form a rule, the less interpretable and interesting it is for an expert in the problen being solved (Barredo Arrieta et al., 2020; Fernandez et al., 2019). The metric takes values in [0, ∞], therefore rules involving higher achievement values may be more useful to the user (see Subsection 2.2).

The objective interestingness is used to consider the potential interest of the rules to the user. Consequently, we have selected the measure CF (Shortliffe & Buchanan, 1975), which provides values in the interval [-1, 1] with negative values representing negative correlation, zero representing independence, and positive values representing positive correlation (see Subsection 2.2). In addition, to face the problems of classic support and confidence measures, very strong rules are the only ones considered as relevant (Berzal et al., 2002). A rule $X \rightarrow Y$ will be considered to be very strong if it fulfils the following conditions: $support(X \rightarrow Y) >$ minSup, $support(\neg Y \rightarrow \neg X) >$ minSup, and $CF(X \rightarrow Y) > 0$. That is, finding evidence of the rule and its opposite in the data, since both rules are equivalent, reinforces the belief that the rule is relevant.

Each rule of the population is evaluated making use of the non-dominance criteria, according to which a rule r_1 dominates another r_2 if for all objectives r_1 has a value that is equal to or better than r_2 and r_1 has a value better than r_2 in at least one of the objectives. Thus, the current poputation is evaluated in the following manner. First, all non-dominated rules are ranked as 1 and are marked (Pareto front rules). Then, all the non-dominated rules, considering only the unmarked rules, are ranked as 2 and are marked from the current population. This process is iteratively applied until all rules from current population are ranked. Finally, the rules with the same rank are also

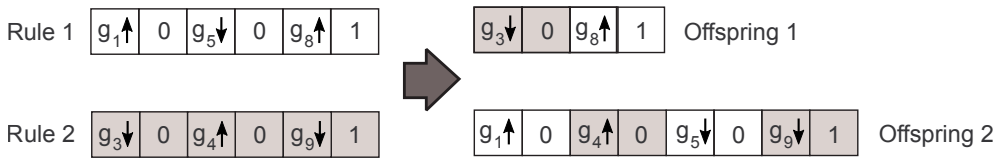

Figure 2: Crossover operator illustration.

ranked taking the crowding measure into account (for a detailed description, see (Deb et al., 2002)).

3.3. Genetic operators

The genetic operators are crossover and mutation. Crossover is performed by random selection of two parents applying binary tournament. Then, 2 offspring are generated interchanging the genes of the selected parents randomly. Crossover operator is illustrated in Figure 2.

Mutation randomly selects a gene from the chromosome and modifies its two parts (*ev* and *ac*). The value of *ev* is updated selecting one event from sequence dataset at random. Then, a value generated at random from the set {0,1} is obtained for *ac*.

In addition, a repairing operator is applied after the crossover and mutation operators to check that the rule antecedents and consequents are consistent. If a rule has no antecedent or consequent, a random event is selected from the events that were not included in the rule. For consequents with two or more events, the consequent is formed with one of the events selected at random and the remaining ones are relocated into the antecedent.

3.4. External Population and Restarting Mechanism

In this proposal we have improved the evolutionary model of NSGA-II by incorporating a restarting mechanism and an EP to encourage diversity in the search process, preserving all the extracted rules. The EP is a repository that allows us to keep all the non-dominated sequential rules obtained. The size of the EP is not limited, which allows us to obtain a greater number of sequential rules from the Pareto front, regardless of the population size of the evolutionary model, which makes it possible to reduce its size to improve the convergence of the search process and to use a fixed size for all datasets (it provides independence with respect to a particular dataset). When the process ends up, the EP usually contains a reduced number of sequential

rules due to the fact that it only collects rules from the Pareto front rules and redundant rules are eliminated to avoid overlaps.

The restarting mechanism introduces diversity in the population, allowing us to avoid premature convergence and ensuring the desired diversity to promote an appropriate exploration. This mechanism runs when the current population size is still less than an $\alpha\%$ of the whole size (where α is commonly fixed as 5%). If so, we update the EP, mark the sequences from dataset in which they occur, and the population is re-initialized (see Subsection 3.1).

3.5. Multiobjective evolutionary model

In this section, we briefly describe the evolutionary model applied in our proposal. The population is initialized, making use of the proposed process to represent information from the complete sequence dataset. Then, a new candidate population is constructed by using selection from the current one, and the described genetic operators, i.e., crossover and mutation. Finally, the next population can be obtained by elitism selecting the best rules from the current and the candidate populations. Then, the non-dominated rules of the current population extend the EP, removing the redundant ones in order to avoid overlaps. Finally, if the current population percentage of new rules is less than $\alpha\%$ then the restarting mechanism is applied, updating the EP and re-initializing the current population. This process is repeated until a stopping condition is satisfied.

3.6. Pseudocode of HAUS-rules

The following flowchart summarizes HAUS-rules.

Input: number of chromosomes in the population ($nChrom$), total number of evaluations ($nEval$), mutation probability (Mut_p), restarting threshold (α).

Output: HAUSRs.

1. Initialization.
   (a) The initial population (P_t) is initialized.
   (b) P_t is evaluated.
   (c) The EP is initialized.
2. Candidate population (CP_t) is generated as follows.

(a) Selection of two parents (x_1 and x_2) by binary tournament selection.

(b) Crossover operator is applied on the individuals x_1 and x_2 to generate two offspring.

(c) Mutation operator is applied with a probability Mut_p on the two generated offspring.

(d) Repairing operator is applied.

(e) The offspring are evaluated.

3. The next population (P_{t+1}) is obtained as follows.

(a) P_t and CP_t are merged.

(b) All non-dominated chormosomes are ranked (rank 1, rank 2,...) and the chromosomes with the same rank are also ranked taking crowding-distance into account.

(c) The best $nChrom$ chromosomes are selected from the merged population.

(d) The non-dominated chromosomes of P_{t+1} extend the EP, removing the redundant ones.

4. If P_{t+1} percentage of new chromosomes is less than $\alpha\%$, the restarting mechanism is applied, updating the EP and re-initializing the current population.

5. If $nEval$ is not achieved, go to 2.

6. The HAUSRs from EP are returned.

4. Experimentation

In this sectión, we illustrate the performance of our proposal. To accomplish this, we establish an experimental analysis based on two stages:

- We compare the performance of our proposal with other approaches of the literature (Section 4.1).
- We analyze the biological relevance the rules generated by the proposed algorithm on in vivo human temporal microarray data using external biological resources (Section 4.2).

Table 5: Summary of the main characteristics of the datasets.

Database	*Sequences*	*Items*	*AvLength*
Bible	36,369	13,905	21.6
BMS	77,512	3,340	4.62
FIFA	20,450	2,990	34.74
Kosarak	10,000	10,094	8.14
SIGN	730	267	51.997
Synthetic 1	1,000	250	45.6
Synthetic 2	2,000	250	49.8
Synthetic 3	5,000	300	50.2
Synthetic 4	7,000	300	49.7
Synthetic 5	10,000	300	52.01

4.1. Analysis of the performance increase with respect to another approach

We have assessed the performance of our proposal using 10 sequence datasets with utility values to assess the performance of our proposal. The quantity of sequences in these datasets is in the range of [730, 77512] and the quantity of events is in the range of [250, 13905]. Table 5 shows the quantity of sequences ("Sequences"), the quantity of different items ("Items"), and the average sequence length ("AvLength") for each dataset. The first 5 datasets (Bible, BMS, FIFA, Kosarak and SIGN) are available in the repository SPMF (Fournier-Viger et al., 2014) from which they can be downloaded [1], and the rest of the datasets were generated using the tools available in SPMF. To develop the different experiments, the external utility for each item has been generated using a log-normal distribution (Tseng et al., 2013). For each dataset, our proposal has been run with 5 different seeds. The values given in the tables represent the mean results average for each one of the metrics by dataset and method.

In these experiments, we have compared our proposal with the well-known algorithm CMRules (Fournier-Viger et al., 2012) and with the algorithm HUSRM (Zida et al., 2015) to illustrate the relevance of the utility of the items and other interestingness measures in the rules extraction process, and to show the effect of the length of the rules on their utility (both algorithms are available in the open-source data mining library SPMF (Fournier-

[1] Available at https://www.philippe-fournier-viger.com/spmf/

Table 6: Parameters used for the comparison.

Algorithms	*Parameters*
CMRules	*minConf = 0.7, minSup = 0.01, minUti = Adapted*
HUSRM	*minConf = 0.7, maxLength = 15, minUti = Adapted*
HAUS-rules	*PopSize = 200,* N_{eval}*=150,000,* P_{mut}*= 0.1,* α *= 5*

Viger et al., 2014)). CMRules is an algorithm based on the Apriori algorithm (Srikant & Agrawal, 1996) that mines sequential rules (with any particular order defined within antecedents, or within consequents) with support and confidence values greater than or equal to a $minSup$ and a $minConf$, respectively, without taking into account the utility of the items. HUSRM is a one-phase algorithm that makes use of a data structure called utility-table and several optimization methods to eliminate unpromising items and efficiently extract HUSRs with confidence and utility values greater than or equal to a $minConf$ and a $minUti$, respectively, without taking into account either the effect of length on utility or the support of the rules in the extraction process. Table 6 shows the parameters used for the analyzed algorithms. The parameters considered for the proposed method in this contribution have been experimentally tuned and we recommend them as standard parameters to work well in most cases. We have selected the values for the rest of the algorithms (with comparative purposes) following the standards and recommendations given by each of their respective authors in the original proposal papers. Notice that, we have adapted the value of $minUti$ for each dataset in the HUSRM and CMRules algorithms because they cannot run in all datasets with a fixed value for this threshold. In addition, in the CMRules algorithm, we have used 0.005 as $minSup$ in Kosarak to be able to extract rules and 0.4 as $minSup$ in SIGN and all synthetic datasets to avoid scalability problems.

Table 7 shows the results average and, in parentheses, standard deviation (σ). Moreover, we show the mean of the quantity of rules obtained $\#R$ and the mean of the quantity of items/events forming the rules Av_{Amp}. Finally, we also show the mean of the interestingness metrics support, confidence, Lift, CF and yule'sQ as Av_{Sup}, Av_{Conf}, Av_{Lift}, Av_{CF}, and $Av_{Yule'sQ}$, respectively. Table 8 shows the results obtained for the utility measures, where $aveUti$ and $CF * aveUti$ are, respectively, the average value of the Average-Utility and the product of CF and Average-Utility of the extracted rules (see Subsection 2.2). Notice that the results obtained on

Table 7: Results obtained for the interestingness measures in all datasets.

Algorithm	$\#R(\sigma)$	$Av_{Amp}(\sigma)$	$Av_{Sup}(\sigma)$	$Av_{Conf}(\sigma)$	$Av_{Lift}(\sigma)$	$Av_{CF}(\sigma)$	$Av_{Yule'sQ}(\sigma)$
CMRules	348.22 (349.54)	**3.27** (0.51)	**0.38** (0.27)	0.83 (0.03)	3.52 (4.31)	0.20 (0.38)	-0.24 (0.79)
HUSRM	**10,164.30** (16,161.56)	9.15 (2.34)	0.07 (0.13)	0.82 (0.08)	8.74 (19.35)	0.59 (0.27)	0.42 (0.48)
HAUS-rules	15.00 (10.02)	3.32 (0.41)	0.22 (0.14)	**0.89** (0.10)	**8.84** (16.43)	**0.68** (0.12)	**0.73** (0.20)

Table 8: Results obtained for the utility measures in all datasets.

Algorithm	$aveUti(\sigma)$	$CF * aveUti(\sigma)$
CMRules	**28,874.54** (31541.28)	649.88 (3,626.57)
HUSRM	10,233.82 (26,248.40)	5,737.47 (15,863.02)
HAUS-rules	23,986.97 (22,207.98)	**17,126.37** (15,853.04)

each dataset can be found on the web page associated with the paper at http://sci2s.ugr.es/HAUS-rules/.

At first glance, given the average results reported in Tables 7 and 8, we can observe how:

- The rules obtained by CMRules present good average values for confidence as the extraction process is guided by the classical measures of support and confidence. However, the average values obtained for other interestingness measures (CF or $Yule'sQ$) are low due to the fact that classical measures do not satisfy some desirable properties, leading to misleading rules that may lead the expert to make wrong decisions when analyzing them. On the other hand, HUSRM allows more specific rule sets to be generated (with an average support of 0.07) due to the fact that this algorithm does not consider a $minSup$, providing rules with average values for confidence equal to or higher than $minCof$ and with higher average values than CMRules for all other measures of interest.

- HUSRM provides a large number of specific rules (including more than 9 items on average) that contain frequent items (which maintain the confidence of the rule) with low utility for the expert, so that the rules have the lowest $aveUtility$ of the analyzed methods. In addition, many of these rules are redundant, in which the only difference between rules are frequent items with low utility. The large number of items involved in the rules and the items with low utility make these rules more difficult to understand and may lead to wrong conclusions when

analyzed. On the other hand, CMRules avoids generating very specific rules by making use of $minSup$, which allows the rules to present a good $aveUtility$. However, the average values of the rules for the interestingness measures are low, so that when the information from both types of measures (interest and utility) is combined, they present the lowest values of the analyzed algorithms.

- HAUS-rules allow to obtain reduced sets of HAUSRs, which present average values that are similar to or even better than the remaining methods for all their interestingness measures and the utility measure $CF * aveUti$. In addition, as these rules are comprised of a decreased quantity of items, they should be easy to understand.

To compare these results, we have used nonparametric tests for multiple comparison to find the best algorithm (Garcia et al., 2009)[2]. The average values that the studied methods are obtaining have been analyzed by application of Statistical tests on the confidence, CF, Yule'sQ, and $CF * aveUti$ measures by previously extrapolating the values to the interval [0, 1]. Notice that we have not considered the measure Lift in this analysis because it represents positive dependence when its value is > 1 (which is the case in all the methods analyzed), and higher values for this measure do not imply a greater interest to the user. In addition, we consider only the utility measure $CF * aveUti$ to take the interest of the rules into account when we analyze their utility.

In spite of finding meaningful inequalities among the studied methods, we apply the Friedman test (Friedman, 1937). Shaffer test (Shaffer, 1986) has been considered once the equality hypothesis has been rejected by Friedman in order to accomplish all the method's pairwise comparisons. Adjusted p-values (APVs) have been computed for each observation in order to ensure the smallest degree of significance for rejecting the equality hypothesis.

Table 9 shows the statistical output for the comparisons. Given the reported APV values, there are significant differences among the observed results with a level of significance of at least 0.1. Consequently, taking these results into account, we can deduce that our proposal was the best performing algorithm among those compared.

Several experiments have been carried out to analyze the scalability of

[2]http://sci2s.ugr.es/sicidm/

Table 9: Statistical test results on the measures confidence, CF, Yule'sQ, and $CF*aveUti$.

Algorithm	Ranking $Conf$	Ranking CF	Ranking $Yule'sQ$	Ranking $CF*aveUti$	APV $Conf$	APV CF	APV $Yule'sQ$	APV $CF*aveUti$
CMRules	2.25	2.75	2.55	2.57	0.1	0.005	0.015	0.012
HUSRM	2.35	1.9	2.13	2.15	0.095	0.057	0.059	0.057
HAUS-rules	**1.4**	**1.35**	**1.3**	**1.3**	-	-	-	-

Table 10: Runtime and memory usage of the algorithms with the increase of the number of sequences in the dataset Synthetic 5.

% sequences	Runtime(seconds) CMRules	HUSRM	HAUS-rules	Memory (MB) CMRules	HUSRM	HAUS-rules
20%	131	>24hour	**92**	4,000	838	**642**
40%	279	>24hour	**173**	7,095	2,160	**700**
60%	532	>24hour	**297**	9,214	3,733	**1,647**
80%	707	>24hour	**385**	9,592	3,926	**1,950**
100%	1,033	>24hour	**579**	10,200	4,180	**1,986**

the algorithms in the dataset Synthetic 5, which presents a good number of sequences and a high average number of items per sequence. In order to perform these experiments, we have used an Intel Core i7, 2.90 GHz CPU with 16GB of memory and running Linux to perform all of the experiments. Table 10 shows the runtime and memory usage by the analyzed algorithms when the number of sequences increases in the dataset Synthetic 5. In addition, Table 11 shows the runtime and memory usage by the analyzed algorithms in all the datasets.

Analyzing the results shown in Table 10 and 11, we can see how the runtime and memory usage scales quite linearly when the number of sequences in the dataset increases. The runtimes of the CMRules algorithm are reduced in all the datasets but a large amount of memory is needed. Therefore, the *minSup* was increased in several datasets to be able to perform the experiments (as noted above in this section). The HUSRM algorithm makes use of a reduced amount of memory for mining the rules but makes use of a large amount of time in most of the datasets. Finally, the HAUS-rules algorithm allows us to extract rules with a good trade-off between runtime and memory usage in all the datasets.

Table 11: Runtime and memory usage of the algorithms in all the datasets.

Dataset	Runtime(seconds)			Memory (MB)		
	CMRules	HUSRM	HAUS-rules	CMRules	HUSRM	HAUS-rules
Bible	22	58	4,717	1,443	3,900	2,211
BMS	1	>7hours	2,091	39	846	1,500
FIFA	-	>24hours	6,644	-	4,200	2,208
Kosarak	1	>24hours	5,601	102	2,200	800
SIGN	7	1,241	30	320	658	528
Synthetic 1	86	>24hours	48	2,950	690	552
Synthetic 2	134	>24hours	92	3,752	857	648
Synthetic 3	428	>24hours	197	6,457	3,400	1,524
Synthetic 4	762	>24hours	377	7,892	4,099	1,849
Synthetic 5	1,033	>24hours	579	10,200	4,180	1,950

4.2. Analysis of the biological relevance of the rules obtained by our proposal in the biological problem dataset

In this subsection, we analyze the biological relevance of a group of rules mined by our method in a human molecular dataset from 57 obese individuals part of a weight-loss dietary trial (see Subsection 2.1). This dataset consists of gene expression data (three temporal records available) recorded during the development of 2 dietary programs (LCD and VLCD). From this data w e have mined HAUSRs that could illustrate the weight-loss induced molecular responses of adipose tissue in obesity. During this approach, the internal utility per item was computed according to the number of occurrences of each item (gene) in the Phenopedia database (Yu et al., 2009), an updated and online resource that gathers all human genetic associations reported to date in the NCBI. The external utility per item was otherwise assessed by means of the absolute fold change of expression reported for a particular item at each time period (this dataset can be found on the web page associated with the paper at http://sci2s.ugr.es/HAUS-rules/). Thanks to these two criteria, we were able to guide the knowledge-extraction process not only by employing expert knowledge, but also by considering the inherent particularities of the experimental design under study. By forcing the algorithm to focus only on the top loci that constitute the genetics architecture of obesity, our proposal omits the production of noisy rules and interfering gene expression patterns susceptible of being tested and validated in wet-lab experiments. After evaluating resulting HAUSRs, a group of experts also found them to be a group of hypotheses with high utility and comprehensibility.

Table 12: Selection of LCD sequential rules with biological relevance.

Rule	MF	BP	SP	CC	TF
$R_1 : (8062821/\downarrow) \rightarrow (7896214/\downarrow)$	6.00	6.00	1.20	6.00	0
$R_2 : (7935776/\downarrow) \rightarrow (8019392/FASN\uparrow)$	6.00	6.00	1.20	6.00	0
$R_3 : (8062821/\downarrow) \rightarrow (7894918/\downarrow)$	6.00	6.00	1.20	6.00	0
$R_4 : (8105084/C7\uparrow) \rightarrow (7940762/LGALS12\uparrow)$	6.00	1.03	6.00	1.03	0
$R_5 : (7935776/\downarrow) \rightarrow (8071036/S100B\uparrow)$	6.00	6.00	1.20	6.00	0
$R_6 : (7935776/\downarrow)$ *and* $(7912692/HSPB7\downarrow) \rightarrow (8019392/FASN\uparrow)$	6.00	6.00	1.20	6.00	0
Average (Standard Deviation)	6.00 (0.0)	5.17 (2.03)	2.00 (1.96)	5.17 (2.03)	0 (0.0)

The evaluation of results by experts was accomplished according to two criteria: 1) Evaluation of the biological quality metrics calculated per mined rule (Molecular Function (MF), Biological Process (BP), Signalling Pathway (SP), Cellular Compartment (CC), and Transcription Factor (TF)), which are all based on external biological information, such as functional and pathway annotation databases (Anguita-Ruiz et al., 2020)), and 2) Contrasting them with the scientific literature. Through this strategy, HAUSRs were filtered according to their biological meaning and contextualized within the obesity background.

Four of our biological metrics (MF, BP, SP and CC) represent rankings calculated from the identical matches (between the items/genes of the antecedent and consequent of the rule) for GO and KEGG biological terms annotated at gene level (Han et al., 2017; Kanehisa et al., 2011). Therefore in these biological measures, a lower resulting value will indicate that the rule is potentially a causal and biologically relevant gene interaction. Range values for these biological measures vary between 6 and 1, with values near to 1 indicating the highest rates of biological concordance between the elements of a rule. Otherwise, the TF measure may take up to four possible values (0, 1, 2 or 3), which are assigned according to the TF-target gene regulatory information hosted in the TRRUST database. For this measure, a 0 value represents the absence of a TF-target gene relationship between the items of the antecedent and consequent of the rule, while higher values (with a maximum of 4) indicate stronger evidence for the existence of a TF-target relationship.

From the application of our proposal to the biological dataset, 6 HAUSRs were mined in the LCD and 9 in the VLCD. Tables 12 and 13 show the extracted HAUSRs and the results obtained for the biological quality mea-

Table 13: Selection of VLCD sequential rules with biological relevance.

Rule	*MF*	*BP*	*SP*	*CC*	*TF*
R_1 : (7896700/ $\uparrow$) $\rightarrow$ (7902623/$DNASE2B\downarrow$)	6.00	6.00	6.00	1.32	0
R_2 : (8091723/$RARRES1\uparrow$) $\rightarrow$ (8093053/$TFRC\downarrow$)	1.62	1.62	1.20	1.62	0
R_3 : (8034304/$ACP5\uparrow$) $\rightarrow$ (8093053/$TFRC\downarrow$)	1.43	1.43	6.00	1.43	0
R_4 : (8046124/$DHRS9\uparrow$) $\rightarrow$ (8046124/$DHRS9\downarrow$)	1.00	1.00	1.20	1.00	0
R_5 : (8091723/$RARRES1\uparrow$) $\rightarrow$ (7923562/$CHIT1\downarrow$)	1.81	1.81	1.20	1.81	0
R_6 : (7935776/ $\downarrow$) *and* (7929816/$SCD\downarrow$) $\rightarrow$ (8165684/ $\uparrow$)	6.00	6.00	1.20	6.00	0
R_7 : (7936322/$GPAM\downarrow$) *and* (8046124/$DHRS9\uparrow$) $\rightarrow$ (8046124/$DHRS9\downarrow$)	1.62	1.62	1.20	1.62	0
R_8 : (7935776/ $\downarrow$) *and* (7929816/$SCD\downarrow$) *and* (8033818/$OLFM2\downarrow$) $\rightarrow$ (8165684/ $\uparrow$)	6.00	6.00	1.20	6.00	0
R_9 : (7936322/$GPAM\downarrow$) *and* (8092177/$NCEH1\uparrow$) *and* (8126784/$PLA2G7\uparrow$) $\rightarrow$ (8000184/$IGSF6\downarrow$)	1.93	1.93	1.20	1.93	0
Average (Standard Deviation)	3.05 (2.23)	3.05 (2.23)	2.27 (2.12)	2.52 (1.99)	0 (0.0)

sures. In both groups, the rules obtained involved genes participating in molecular processes previously reported as part of the weight-loss induced adipose tissue response (e.g. proliferation and inflammatory pathways or lipid metabolism) (Vink et al., 2016). According to group mean quality measures values, better values were derived in the VLCD than in the LCD, given the higher molecular impact elicited by this dietary program in the adipose tissue. All resulting rules presented good rates in at least one of the biological quality measures, except for the TF measure. The absence of TF scores other than zero among the sequential rules discovered could be explained by the fact that no transcription factors are participating in the gene expression patterns identified, which does not imply that discovered sequences are not of biological interest.

In order to further assess the biological utility of each set of sequential rules, obesity-field experts investigated the whole set of HAUSRs, contrasting them with the scientific literature and visually inspecting two graphical representations in circular plots (Figures 3 and 4 for LCD and VLCD respectively), in which node names refer to probe/gene ids, edges represent $X \rightarrow Y$ connections between gene/probes, vertex diameter refers to the value of CF of each rule, vertex colour represents the type of gene expression change (dark blue for up-regulations and light blue for down-regulations), edge colour represents the value for the BP quality measure (i.e. best values are represented with darker colours), and edge width represents values for the SP measure (i.e. top values are represented with wider edges).

From the set of sequential rules found in the LCD group, the rule R_4 is highlighted, as it gave the best score for the biological quality metric BP.

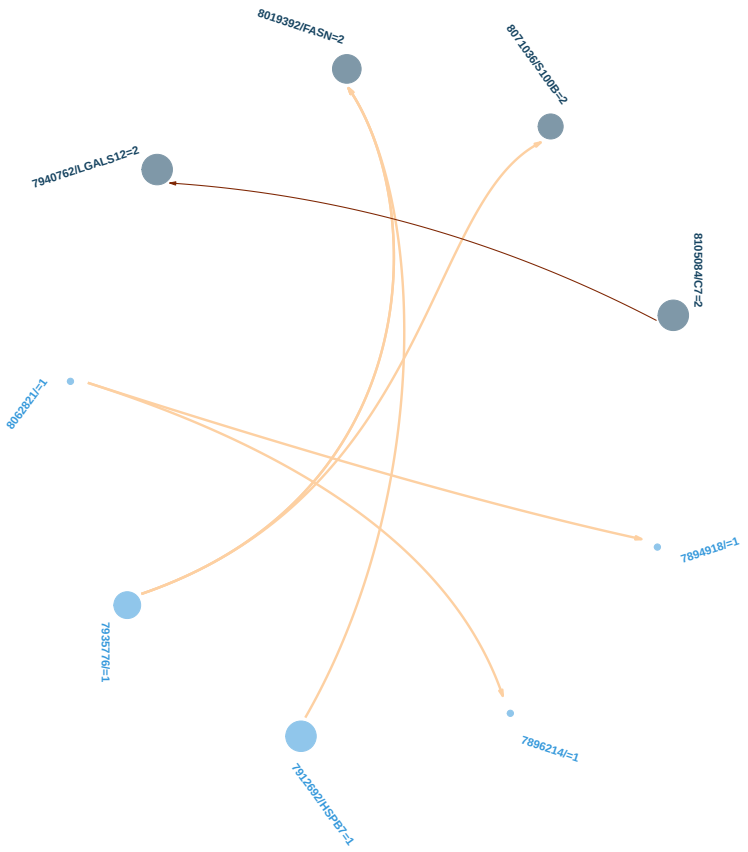

Figure 3: Hierarchical edge bundling representation for LCD sequential rules.

This indicates that both genes participate in the same biological process/es. Following this particular pattern, we can see how an upregulation of the gene C7 (or complement C7), which is involved in the immune response Lectin induced complement pathway, is followed by or caused by a posterior upregulation of the expression of LGALS12. Interestingly, LGALS12 (or galectin-12) is a lectin human protein, expressed in the adipose tissue and involved in the regulation of lipid degradation (lipolysis) (Yang et al., 2012). By its function, LGALS12 has been suggested to be a useful target for the treatment of obesity-related metabolic conditions, such as insulin resistance, metabolic syndrome, and type 2 diabetes (Yang et al., 2012). Although further validation for this interaction is needed, it reveals a plausible mechanism

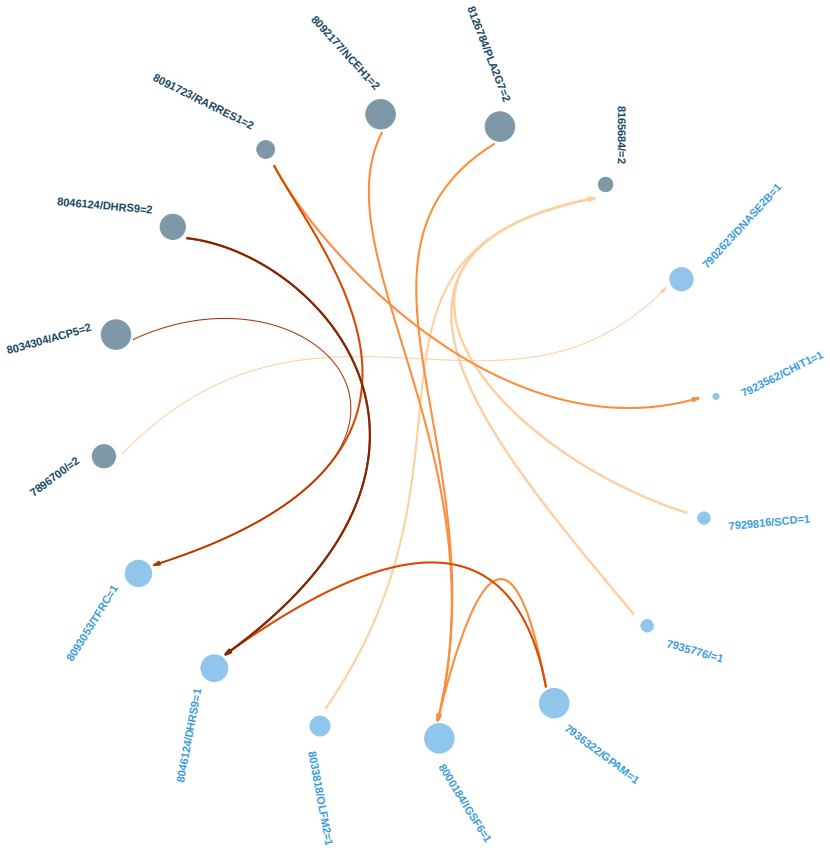

Figure 4: Hierarchical edge bundling representation for VLCD sequential rules.

by which lectins elicits its negative effects in obesity and the immune system.

Rule R_4 stands out from the set of VLCD sequential rules, as it gives the highest score for the biological quality metric BP. With this interaction pattern, we observe how, although certain genes experiment an upregulation in response to weight-loss, their expression levels return to a baseline value at the moment a normal-calorie diet is resumed during the weight-stabilization phase (i.e. negative feedback loop regulation of their own expression). Interestingly, this kind of phenomenon could argue the existence of fast transient dynamic molecular changes in human tissue in response to interventions. Another interesting pattern identified in the VLCD group is the one identified between the genes ACP5 and TFRC, whose orchestrated regulation and joint participation in molecular networks have been previously described in obesity (Marrades et al., 2011).

Considering all this, we can state that the method was able to generate comprehensible rules with strong biological meaning and of great interest for biologists. Thus, the application of this method could be easily extended to other human longitudinal interventions counting on gene expression data.

5. Conclusions

In this paper, we have proposed HAUS-rules, a new MOEA for mining HAUSRs from sequence datasets. To do so, we have extended the well-known and widely applied NSGA-II in order to perform an evolutionary learning and condition selection for each sequential rule, while introducing an external population and a restarting process to encourage diversity in the search process, preserving all the rules obtained. In addition, this proposal maximizes two objectives -utility and interestingness- in order to obtain rules that are interesting, easy to understand and with a high average-utility value.

The results obtained from the ten datasets show that our proposal allows us to mine a reduced set of HAUSRs with high interestingness and utility values in all datasets. In addition, these rules involve few items/events, making it easier to comprehend from a user's perspective. The rules extracted from human long-term intervention data present a high biological relevance according to their values for several biological measures, which represent plausibly mechanisms that illustrate (although further validation of these interactions is needed) weight-loss induced gene regulatory responses of adipose tissue in obesity.

6. Acknowledgements

This work was supported by the ERDF/Regional Government of Andalusia/Ministry of Economic Transformation, Industry, Knowledge and Universities (grant number P18-RT-2248) and the ERDF/Health Institute Carlos III/Spanish Ministry of Science, Innovation and Universities (grant number PI20/00711).

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD* (pp. 207–216). Washington D.C., USA.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *11th International Conference on Data Engineering (ICDE 1995)* (pp. 3–14). Washington, DC, USA.

Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C. M., & Alcalá-Fdez, J. (2020). eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLOS Computational Biology*, *16*, 1–34.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82 – 115.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research*, *41*, D991–D995.

Berzal, F., Blanco, I., Sánchez, D., & Vila, M. A. (2002). Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, *6*, 221–235.

Brin, S., Motwani, R., Ullman, J., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*, *26*, 255–264.

Chamazi, M., & Motameni, H. (2019). Finding suitable membership functions for fuzzy temporal mining problems using fuzzy temporal bees method. *Soft Computing*, *23*, 3501–3518.

Deb, K. (2015). Multi-objective evolutionary algorithms. In J. Kacprzyk, & W. Pedrycz (Eds.), *Springer Handbook of Computational Intelligence* (pp. 995–1015). Berlin, Heidelberg: Springer Berlin Heidelberg.

Deb, K., Agrawal, S., Pratab, A., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions Evolutionary Computation*, *6*, 182–197.

Dinh, D., Le, B., Fournier-Viger, P., & Huynh, V. (2018). An efficient algorithm for mining periodic high-utility sequential patterns. *Applied Intelligence*, *48*, 4694–4714.

Diop, L., Diop, C. T., Giacometti, A., Li, D., & Soulet, A. (2020). Sequential pattern sampling with norm-based utility. *Knowledge and Information Systems*, *62*, 2029–2065.

Fernandez, A., Herrera, F., Cordon, O., Jose del Jesus, M., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine*, *14*, 69–81.

Fournier-Viger, P., Faghihi, U., Nkambou, R., & Nguifo, E. M. (2012). CM-Rules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, *25*, 63–76.

Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., & Tseng, V. S. (2014). SPMF: A java open-source pattern mining library. *Journal of Machine Learning Research*, *15*, 3569–3573.

Fournier-Viger, P., Lin, J.-W., Kiran, R., Koh, Y., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, *1*, 54–77.

Fournier-Viger, P., Wu, C., Tseng, V. S., Cao, L., & Nkambou, R. (2015). Mining partially-ordered sequential rules common to multiple sequences. *IEEE Transactions on Knowledge and Data Engineering*, *27*, 2203–2216.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Statist. Assoc.*, *32*, 675–701.

Gallo, C. A., Cecchini, R. L., Carballido, J. A., Micheletto, S., & Ponzoni, I. (2016). Discretization of gene expression data revised. *Briefings in Bioinformatics*, *17*, 758–770.

Gan, W., Lin, J. C.-W., Zhang, J., Fournier-Viger, P., Chao, H.-C., & Yu, P. S. (2021). Fast utility mining on sequence data. *IEEE Transactions on Cybernetics*, *51*, 487–500.

Garcia, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non-parametric tests for analyzing the evolutionary algorithmsbehaviour: A case study on the cec2005 special session on real parameter optimization. *Journal Heuristics*, *15*, 617–644.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, *38*, 1–32.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H., & Lee, I. (2017). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*, D380–D386.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. (2nd ed.). USA: Morgan Kaufmann.

Hong, T.-P., Lee, C.-H., & Wang, S.-L. (2011). Effective utility mining with the measure of average utility. *Expert Systems with Applications*, *38*, 8259 – 8265.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., & Hobbs, B. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, *31*, e15.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*, D109–D114.

Kilgore, P. C. S. R., Korneeva, N., Arnold, T. C., Trutschl, M., & Cvek, U. (2019). Gatewaynet: a form of sequential rule mining. *BMC Medical Informatics and Decision Making*, *19:87*, 1–13.

Liang, Y., & Kelemen, A. (2017). Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. *Briefings in Bioinformatics*, *19*, 1051–1068.

Liu, Z., Xue, Y., Li, M., Ma, B., Zhang, M., Chen, X., & Hu, X. (2017). Discovery of deep order-preserving submatrix in dna microarray data based on sequential pattern mining. *International Journal of Data Mining and Bioinformatics*, *17*, 217–237.

Marrades, M., González-Muniesa, P., Arteta, D., Alfredo Martínez, J., & Moreno-Aliaga, M. (2011). Galectin-12: A protein associated with lipid droplets that regulates lipid metabolism and energy balance. *Journal of Physiology and Biochemistry*, *67*, 15–26.

Matthews, S. G., Gongora, M. A., & Hopgood, A. A. (2013). Evolutionary algorithms and fuzzy sets for discovering temporal rules. *International Journal of Applied Mathematics and Computer Science*, *23*, 855–868.

Nasu, M., Shimamura, K., Esumi, S., & Tamamaki, N. (2020). Sequential pattern of sublayer formation in the paleocortex and neocortex. *Medical Molecular Morphology*, *53*, 168–176.

Nguyen, D., Luo, W., Phung, D., & Venkatesh, S. (2018). LTARM: A novel temporal association rule mining method to understand toxicities in a routine cancer treatment. *Knowledge-Based Systems*, *161*, 313–328.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M.-C. (2001). PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering* (pp. 215–224). Heidelberg, Germany.

Ramaswamy, S., Mahajan, S., & Silberschatz, A. (1998). On the discovery of interesting patterns in association rules. In *24rd International Conference on Very Large Data Bases* (pp. 368–379). California, USA.

Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826–831.

Sheskin, D. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K.: Chapman & Hall/CRC.

Shortliffe, E., & Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*, 351–379.

Srikant, R., & Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In *International Conference on Management of data (ACM SIGMOD 1996)* (pp. 1–12).

Srivastava, G., Singh, A., & Mallipeddi, R. (2021). NSGA-II with objective-specific variation operators for multiobjective vehicle routing problem with time windows. *Expert Systems with Applications*, *176*, 114779.

Tan, P., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *8th International Conference on Knowledge Discovery and Data Mining (KDD 2002)* (pp. 32–41). Edmonton, Canada.

Truong, T., Duong, H., Le, B., & Fournier-Viger, P. (2019). Efficient vertical mining of high average-utility itemsets based on novel upper-bounds. *IEEE Transactions on Knowledge and Data Engineering*, *31*, 301–314.

Tseng, V. S., Shie, B., Wu, C., & Yu, P. S. (2013). Efficient algorithms for mining high utility itemsets from transactional databases. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 1772–1786.

Vink, R., Roumans, N., Fazelzadeh, P., Tareen, S., Boekschoten, M., van Baak, M., & Mariman, E. (2017). Adipose tissue gene expression is differentially regulated with different rates of weight loss in overweight and obese humans. *International journal of obesity*, *41*, 309–316.

Vink, R. G., Roumans, N. J. T., Arkenbosch, L. A. J., Mariman, E. C. M., & van Baak, M. A. (2016). The effect of rate of weight loss on long-term weight regain in adults with overweight and obesity. *Obesity*, *24*, 321–327.

Wu, J. M., Lin, J. C., Pirouz, M., & Fournier-Viger, P. (2018). Tub-haupm: Tighter upper bound for mining high average-utility patterns. *IEEE Access*, *6*, 18655–18669.

Yang, J., Wang, W., Yu, P. S., & Han, J. (2002). Mining long sequential patterns in a noisy environment. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data* (pp. 406–417). New York, NY, USA.

Yang, R., Havel, P., & Liu, F. (2012). Galectin-12: A protein associated with lipid droplets that regulates lipid metabolism and energy balance. *Adipocyte*, *1*, 96–100.

Yu, W., Clyne, M., Khoury, M. J., & Gwinn, M. (2009). Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics*, *26*, 145–146.

Yu, X., & Gen, M. (2012). *Introduction to Evolutionary Algorithms*. Springer Publishing Company, Incorporated.

Zhang, C., Du, Z., Gan, W., & Yu, P. S. (2021). Tkus: Mining top-k high utility sequential patterns. *Information Sciences*, *570*, 342–359.

Zhang, C., Han, M., Sun, R., Du, S., & Shen, M. (2020). A survey of key technologies for high utility patterns mining. *IEEE Access*, *8*, 55798–55814.

Zhou, H., & Hirasawa, K. (2019). Evolving temporal association rules in recommender system. *Neural Computing and Applications*, *31*, 2605–2619.

Zida, S., Fournier-Viger, P., Wu, C.-W., Lin, J. C.-W., & Tseng, V. S. (2015). Efficient mining of high-utility sequential rules. In P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition* (pp. 157–171). Cham: Springer International Publishing.

Zihayat, M., Davoudi, H., & An, A. (2017). Mining significant high utility gene regulation sequential patterns. *BMC Systems Biology*, *11(Suppl 6)*, 1–88.