Handling Fuzzy Gaps in Sequential Patterns: Application to Health

Sandra Bringay, Anne Laurent, Béatrice Orsetti, Paola Salle, Maguelonne Teisseire

Handling Fuzzy Gaps in Sequential Patterns: Application to Health

Sandra Bringay, Anne Laurent, Béatrice Orsetti, Paola Salle and Maguelonne Teisseire

***Abstract*— Dealing with numerical data for mining novel knowledge is a non trivial task that has received much attention in the last years. However, it is still not easy to handle such data, especially when large volumes of values must be analyzed. In our work, we focus on biological data from DNA chips that biologists study in order to try and discover new gene correlations that could help understanding diseases like breast cancer. In this framework, we consider the values from the DNA microarrays, which convey the behavior of some genes, and we want to discover how these behaviors are correlated. This data are considered as being ordered as numerical values can be sorted. In previous work, rules like $\langle (1 \quad 5)(2) \rangle$ have been discovered, meaning that genes 1 and 5 have the same expression level followed by gene 2 that has a higher expression value. However, such data are very noisy and considering close values as ordered is often false. We thus consider here fuzzy rankings based on a fuzzy partition provided by the experts. Rules can then better characterize how genes are correlated.**

I. Introduction

Health is a major concern for the modern society. Cancer is a leading cause of death and in particular breast cancer, which is the second most common type of cancer and the fifth most common cause of cancer death [Org09]. As deaths from breast cancer worldwide are projected to continue rising, the problem of discovering the genes involved in their development have been intensively addressed by the biomedical community. In this framework, researchers aim at discovering how genes behave in terms of their implication in biological processes and how their expression is regulated and co-regulated.

Formerly impossible because of the costs, collecting large amounts of data is now possible, especially by means of the microarrays. Such microarrays [BCS$^+$05], [HDG05] allow researchers to compare the expression of genes in different tissues, cells or conditions. However, the way to process those data for making a biomedical sense is a big challenge. For such purposes, data mining plays a key role as it allows for discovering previously unknown knowledge from large amounts of data.

In the framework of our study, we have studied test datasets available online[1] [WKZ$^+$05]. The number of microarrays depend on the number of people whose tumor was taken (286 samples). Each microarray reports the expression of 17,816 genes put under several conditions, leading to tables containing $x \times y$ points. Each point describes the expression value ranging from -7.54 to 8.57. Negative values mean that the gene is under expressed, positive values mean that the gene is over expressed, while a null value means that no expression is detected. Such data are thus given by a matrix as shown by Figure 1 that is often converted to colors (usually yellow, green and red) after discretizing the values, as shown on Figure 2. Such discretized values are often used by existing approaches, for instance in statistical methods [TTC01], [KMC00], in clustering approaches [ESBD98], [PBB08], [MO04], or in association rules mining [JG05].

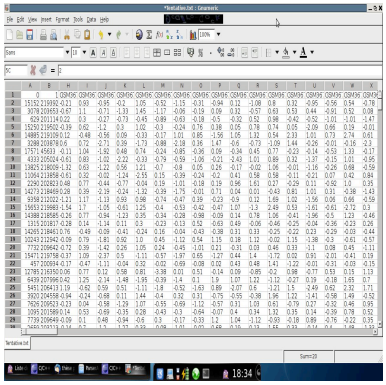

Fig. 1. Gene Expression Values from a DNA Microarray

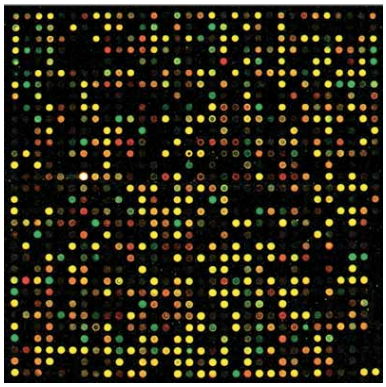

Fig. 2. Gene Expression Colors from a DNA Microarray

In this paper, we focus on such data and we aim at discovering rules to help biologists and medical doctors to analyze how genes interact. In previous work, we have proposed an algorithm DEMON [SBT09] to extract sequential patterns from microarray data. An example of such pattern is *gene 'Gene 1' has an expression lower than genes*

Sandra Bringay, Anne Laurent, Paola Salle and Maguelonne Teisseire are with the LIRMM laboratory - CNRS UMR 5506, Univ. Montpellier 2 (contact email: laurent@lirmm.fr).

Maguelonne Teisseire is with the CEMAGREF Montpellier.

Béatrice Orsetti is with the INSERM.

[1]http://www.ihes.fr/ zinovyev/princmanif2006/

'Gene 2' and 'Gene 3' which expressions are close. In this paper, we focus on how fuzzy logic can help defining rules that are more understandable and actionable for the experts. For instance, the experts are eager to be provided with rules that they can easily interpret in a linguist manner without being obliged to define a crisp partition of the values from the microarrays. For this purpose, we propose to extent classical sequential patterns to obtain rules like $\langle (G1\ G5)_{(very\ overexpressed, 0.8)}(G3)\rangle$, meaning that in the experiment led, *the gene 3 is far much expressed compared to genes 1 and 5, which are expressed in a similar way*. The level to which genes are under or over expressed is defined using a fuzzy partition, leading to the so called *fuzzy gaps*.

We provide below the definitions introduced in order to handle fuzzy gaps, together with the associated algorithms. Experiments reported here show some examples of patterns found by this approach, and could not have been discovered using classical crisp approaches.

II. Basic Definitions

In this section, we present the seminal information related to the data we want to mine, and sequential patterns [AS95], that are the basis of our approach.

A. Biological Data of Breast Cancers

Cancer arises from a change in a single cell, which spreads to one or more organs. This change may be caused by external agents such as tobacco or ionizing radiation and/or inherited genetic factors. Cancers result from mutations, or abnormal changes, in the genes responsible for the regulation of the growth of cells. While the cells of healthy people are normally regenerated by this process, mutations can occur in sick people cells, which will inhibit some genes and increase the action of other ones, resulting in abnormal behaviors. The cells divide without control, producing more cells and forming a tumour. A tumour can be benign (not dangerous) or malignant (potentially dangerous). Breast cancer refers to a malignant tumour that has developed from cells in the breast. Discovering new information about groups of genes implied in breast cancers is challenging. We thus propose to mine for rules describing how the levels of gene expressions are correlated and ordered. This method is based on sequential patterns as they allow us to discover correlations among ordered items. For this purpose, we mine databases that report the level of expressions of genes taken from cells of several patients, as described by the example below.

Example 1: DNA chips report the gene expression levels for some genes taken from the cells of breast tumors. We consider here 4 chips and 5 genes. Table I displays this database, genes having been ordered by expression level.

Chip_ID	Expression Level	Genes
1	-6.2	G2
	-1.8	G1
	-1.3	G5
	2.3	G3
	4.8	G4
2	-5.4	G2
	-2	G1,G5
	2.3	G3
	4.8	G4
3	-5.8	G2
	-3.6	G1
	-3	G5
	2.3	G3
	7.5	G4
4	-4.7	G2
	0	G4
	2.3	G5
	8.57	G3,G1

TABLE I
EXAMPLE OF DATABASE FOR GENES $1, 2, 3, 4, 5$

B. Sequential Patterns

Sequential patterns are often introduced as an extension of association rules in [AS95]. They highlight correlations between database records as well as their temporal relationships. Some generalization were proposed to use fuzzy set theory to handle numeric attributes. In this paper we use sequential patterns with fuzzy gaps to provide the biologists with more precise sequences.

Sequential patterns describe frequent sequences of itemsets. By itemset, we consider a set of items that occur simultaneously in a timestamped database. For instance, when considering a database from a supermarket, the database will describe the items purchased by customers at different dates. Frequent patterns extracted from such database are like $\langle (butter\ \ chips)(chocolate) > sup$ meaning that $sup\%$ of the customers purchased butter together with chips and *then* chocolate.

Definition 1 (Itemset): Let $I = \{i_1, i_2, ..., i_m\}$ be the set of items being considered. An *itemset* is a non-empty set of items $(i_1 i_2 \dots i_k) \subseteq I$.

It should be noted that as an itemset is a set, it is a non-ordered representation.

Definition 2 (Sequence): A *sequence* s is a non-empty ordered list of itemsets, denoted by $\langle it_1 it_2 ... it_p\rangle$. A sequence is said to be an *n-sequence* (or a sequence of size n) if is consists of n items.

Definition 3 (Database): Let $\mathcal{R}$ be a set of object records where each record consists of three pieces of information: an object-id, a record timestamp and a set of items appearing in the record. $\mathcal{R}$ is said to be the database.

In our context, the object-id is an experiment (Chip_ID). Records correspond to the genes, and timestamps correspond to the expression levels, as detailed below.

Example 2: Referring back to Table I, $G1$, $G2$, $G3$, $G4$ and $G5$ are items. $(G1\ G5)$ is a itemset, and $s = \langle (G1\ G5)\ (G3)\rangle$ is a sequence. The database is given by the whole table.

Definition 4 (Subsequence): A sequence $\langle it_1\ \ it_2 ... it_p\rangle$ is a *subsequence* of another one $\langle\ it'_1\ \ it'_2\ \ ...it'_m\rangle$ if there exist integers $l_1 < l_2 < ... < l_p$ such that $it_1 \subseteq it'_{l_1},\ it_2 \subseteq it'_{l_2}, ...,\ it_p \subseteq it'_{l_p}$.

Chip_ID	Gene Expression Sequence
1	$\langle(G2)(G1)(G5)(G3)(G4)\rangle$
2	$\langle(G2)(G1\ G5)(G3)(G4)\rangle$
3	$\langle(G2)(G1)(G5)(G3)(G4)\rangle$
4	$\langle(G2)(G4)(G5)(G3\ G1)\rangle$

TABLE II
EXAMPLE OF SEQUENCE DATABASE FOR GENES $1, 2, 3, 4, 5$

Example 3: Referring back to our previous examples and Table I, the sequence s' = $\langle$(G1) (G3)$\rangle$ is a subsequence of $s =< (G1\ G5)\ (G3)\rangle$ because (G1) $\subseteq$ (G1 G5) and (G3) $\subseteq$ (G3). However, $\langle$(G1) (G5)$\rangle$ is not a subsequence of s.

It should be noted that databases can be displayed as a set of sequences, as shown by Table II. All records from the same object are then grouped together and sorted in increasing order of their timestamp. They are called a data sequence.

Definition 5 (Sequence Support): An object *supports* a sequence s if s is included within the data sequence of this object (s is a subsequence of the data sequence).

The *frequency* of a sequence ($freq(s)$) is defined as the percentage of objects supporting s.

A sequence is said to be frequent according to a minimum frequency value ($minFreq$) specified by the user if the condition $freq(s) \geq minFreq$ holds. The problem of discovering *frequent* sequential patterns is then described as the discovery of all sequences that occur for most than $minsup$ customers, where $minsup$ is said to be the minimal support and is provided by the user. The set of sequential patterns contains all maximal[2] frequent sequences.

In classical approaches, the sequences are discovered whatever the time gap that may occur between two itemsets (e.g., purchases). More recently, time constraints have been introduced in order to manage the inter-relations in a more precise way [SA96]. For instance, such constraints allow for discarding sequences if the gap between two itemsets is greater than a given number of days or, in the opposite, two itemsets are considered as being simultaneous if they occur at close dates. Fuzzy constraints have been introduced by [FLT07], [FLT09]. However these constraints do not allow to consider several gaps between two itemsets.

In this paper, we consider fuzzy intervals to describe these gaps, thus providing an approach to mine fuzzy ordered patterns.

III. FUZZY ORDERED PATTERNS

In this section, we detail our contribution, aiming at providing efficient and relevant methods for mining patterns including fuzzy gaps. For this purpose, we consider a fuzzy partition provided by the expert [JKZ73]. This partition is given as a set of n fuzzy sets $\mathcal{A} = \{A_1, \ldots, A_n\}$ defined over the universe U of the values that can characterize the difference between two gene expression values. This fuzzy partition is defined such as $\forall u \in U$, $\sum_{i=1}^{n} \mu_i(u) = 1$ where μ_i is the membership function of the fuzzy set A_i. Figure 3 describes an example of such a partition. It should be noted that, as asked by the biologists, two genes showing an expression difference lower than $ln(2)$ will be considered as having a similarly expression. $ln(2)$ plays a crucial role in the definition of the fuzzy partition as the normalization of microarrays relies on this particular value.

Based on this partition, it is possible to describe when two values must be considered as being *lightly*, *averagely* or *very* different, and to which extent. These descriptions are used to defined the fuzzy gaps, which will appear in the patterns we introduce in this paper.

A. Mining Sequential Patterns with Fuzzy Gaps

In this section, we detail how we handle fuzzy gaps when mining microarray data. We recall here that rules like $\langle(G1\ G5)_{(very\ different, 0.8)}(G3)\rangle$ are to be extracted, meaning that *genes 1 and 5 have similar expression values, followed by gene 3 that has a very different value*.

Example 4: From Table I, it is clear that, for 3 chips out of 4, the expression of gene $G1$ is lower than the expression of $G3$ and that the expression of gene $G5$ is lower than the expression of $G3$, two frequent sequences would then be discovered: $\langle(G1)(G3)\rangle$ and $\langle(G5)(G3)\rangle$. However, it is impossible to see that $G1$ and $G5$ have similar expressions and to describe how the expressions are separated. In our approach, we thus soften this and, as said above, we mine rules like *Gene 3 is far much expressed compared to genes 1 and 5, which are expressed in a similar way* reported as $\langle(G1\ G5)very\ over\ expressed(G3)\rangle$.

Each rule is associated to its support and to the degrees to which the fuzzy relations (e.g., very close, close, very different) hold.

We provide below the formal approach we propose and the associated definitions.

We recall here that our approach aims at mining databases defined as in Definition 3. Such a database contains objects, each object consisting of a sequence of itemsets. We first define how the crisp gap between two itemsets can be computed.

Definition 6 (Itemset Difference): Let $\mathcal{R}$ be a database, o be a record from $\mathcal{R}$, and it_1 and it_2 be two itemsets from the sequence associated to o. The difference $\delta(it_2, it_1)$ between it_1 and it_2 is defined as the absolute value of absolute value of the difference between the timestamp of the first item of it_2 and the last item of it_1.

Example 5: In Table I, the difference $\delta(it_2, it_1)$ between itemsets $it1 = (G2)$ and $it2 = (G1\ G5)$ equals $|-2-(-5.4)| = 3.4$ for object 2 and $|-1.8-(-6.2)| = 4.4$ for object 1 as -1.8 is the lower expression value for itemset it_2.

[2]in terms of their size

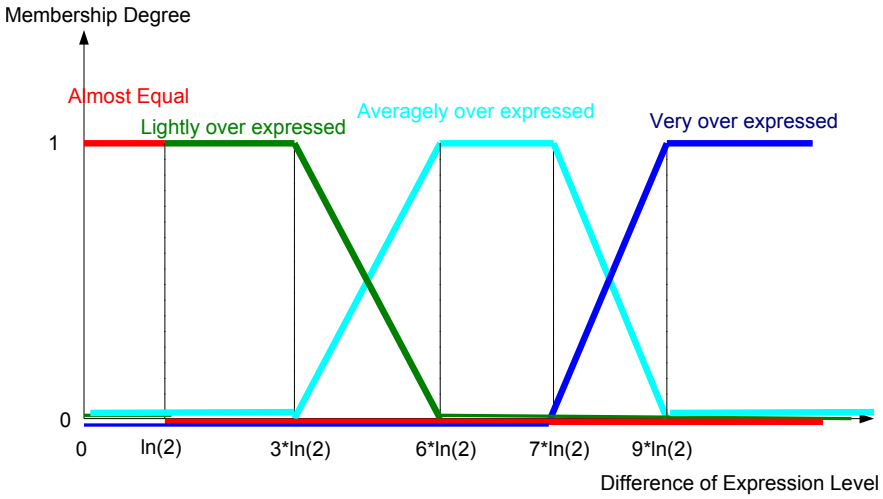

Fig. 3. Fuzzy Partition for Expression the Difference Between Two Gene Expression Values

In our context, calculating the difference amounts to compute the difference between the two expression levels.

Definition 7 (Fuzzy Gap Sequence): Let $\mathcal{A}$ be a set of fuzzy sets building a fuzzy partition, a fuzzy gap sequence s_{FG} is defined as an ordered list of itemsets and fuzzy sets such as $s_{FG} = \langle it_1(A_{i_1})it_2(A_{i_2})...(A_{i_{p-1}})it_p\rangle$ where $\forall i \in [1, p-1]$, $A_i \in \mathcal{A}$.

Definition 8 (Fuzzy Gap Degree): Let s_o be the data sequence associated to an objet o, let it_1 and it_2 be two itemsets and A be a fuzzy subset associated to its membership function μ. A fuzzy gap degree between it_1 and it_2 is a pair (A, d) and we have $\langle it_1(A, d)it_2\rangle$ where $d = \mu(\delta(it_2, it_1))$

Example 6: We consider the object 1 from Table I, and the fuzzy set $averagely\ over\ expressed$ (hereafter avg). We have $\delta(it_2, it_1) = 4.4$ as shown in the previous example. We have the fuzzy gap degree $(G1\ G5)(avg; 1)(G3)$ as $\mu_{avg}(4.4) = 1$.

Considering object 2, we have $(G1\ G5)(avg; 0)(G3)$ as $\delta(it_2, it_1) = 0.3$ and $\mu_{avg}(0.3) = 0$.

Considering object 3, we have $(G1\ G5)(avg; 0.66)(G3)$ as $\delta(it_2, it_1) = 5.3$ and $\mu_{avg}(5.3) = 0.66$.

Considering object 4, the sequence $(G1\ G5)(G3)$ does not hold as $G1$ occurs at a higher level compared to $G5$.

The degree of a data sequence is computed using a t-norm to merge the fuzzy gap degrees. Note that the t-norm $botn$ used is generalized to an n-ary function.

Definition 9 (Fuzzy Gap Sequence Degree): Let s_o be the data sequence associated to an objet o, a fuzzy gap degree-sequence s_{FGD} is defined as $s_{FGD} = \langle it_1(A_{i_1}, d_1)it_2(A_{i_2}, d_2)...(A_{i_{p-1}}, d_{p-1})it_p\rangle$ where (A_i, d_i) are fuzzy gap degrees. The degree of s_o denoted by $F_{s_{FG}}(s_o)$ is then computed as: $F_{s_{FG}}(s_o) = \perp(d_1, ..., d_{p-1})$

Remark. In our approach, the min is considered as the t-norm we use as we do not want to penalize long sequences by over decreasing their degree. Note that every t-norm is associative, and can thus be computed on-the-fly when building the fuzzy gap sequence, thus leading to more performant algorithms.

Finally, the support of a fuzzy gap sequence is computed by summing the membership degrees over the whole database.

Definition 10 (Fuzzy Gap Sequence Support): Let s_{FG} be a fuzzy gap sequence, let $\mathcal{O}$ be the set of objects of the database records, the support of s_{FG} is computed as:

$$Freq(s_{FG}) = \frac{\sum_{o \in \mathcal{O}} \left[F_{s_{FG}}(s_o)\right]}{|\mathcal{O}|}$$

Example 7: In our running example, the sequence $(G1\ G5)(averagely\ over\ expressed)(G3)$ has a support of $\frac{1+0+0.66+0}{4} = \frac{1.66}{4} = 0.415 = 41.5\%$.

Given a database of object records and the associated sequential patterns, the problem of fuzzy gap sequence mining is to find all maximal fuzzy gap sequences which support is greater than a user-defined threshold ($MinFGFreq$).

B. Algorithms

In this section, we detail the algorithms we propose for tackling the problem of mining fuzzy gap sequences.

Sequential patterns are computed using a levelwise approach, meaning that sequences of size k are computed from frequent sequences of size $k-1$. This approach allows us to remain scalable as the databases to be mined can be very large. In the context of fuzzy gap sequences, we aim at describing the gaps between itemsets in frequent sequences in order to display a more valuable information to the experts. For this purpose, each sequential pattern is examined in order to define the best linguistic labels to be associated with the gaps between the itemsets, as reported by Algorithm 1.

Algorithm 1: Fuzzy Gap Sequence Extraction

Data: $\mathcal{R}$, a database of records
$\mathcal{MS}$, the set of sequential patterns from $\mathcal{R}$
$MinFGFreq$, the threshold defined by the user
$\mathcal{A}$ the set of fuzzy sets building a fuzzy partition
Result: $\mathcal{S}_{\mathcal{FG}}$, the set of maximal fuzzy gap sequences
begin
  $\mathcal{S}_{\mathcal{FG}} \leftarrow \varnothing$
  $FuzzyGapSeq \leftarrow$FuzzyGapGen$(\mathcal{MS}, \mathcal{A})$
  /* Generation of all possible fuzzy gap sequences associated to the fuzzy subsets*/
  foreach $s_{FG} \in FuzzyGapSeq$ **do**
    $Freq_{s_{FG}} \leftarrow \varnothing$
  foreach $o \in \mathcal{R}$ **do**
    foreach $s_{FG} \in FuzzyGapSeq$ **do**
      $Freq_{s_{FG}} \leftarrow Freq_{s_{FG}} + F_{s_{FG}}(s_o)$
      /* The frequency is computed by summing the degree of each object data sequence */
  foreach $s_{FG} \in FuzzyGapSeq$ **do**
    if *(*$Freq_{s_{FG}}/|\mathcal{R}| \geq MinFGFreq$ **then**
      $\mathcal{S}_{\mathcal{FG}} \leftarrow \mathcal{S}_{\mathcal{FG}} \cup \{s_{FG}\}$
end

The $FuzzyGapGen$ function (Algorithm 2) generates all combinations of sequences with the possible fuzzy sets for describing the gaps between itemsets.

Algorithm 2: FuzzyGapGen

Data: $\mathcal{MS}$ a set of sequential patterns
$\mathcal{A}$ a set of fuzzy sets
Result: $\mathcal{S}_{\mathcal{FG}}$, the set of all possible fuzzy gap sequences associated to $\mathcal{MS}$ and $\mathcal{A}$
begin
  $\mathcal{S}_{\mathcal{FG}} \leftarrow \varnothing$
  foreach $seq \in \mathcal{MS}$ **do**
    $n \leftarrow sizeof(seq)$;
    /* there are $n-1$ gaps, thus leading to $|card(\mathcal{A})|^{n-1}$ fuzzy gap sequences */
    /* loc_s is the set of sequences for sequence seq */
    $loc_s \leftarrow \{\langle it_1 \rangle\}$
    for *(i=1;* $i < n$ *; i++)* **do**
      foreach $s \in loc_s$ **do**
        $loc_s \leftarrow loc_s - \{s\}$
        /* Concatenate each fuzzy gap sequence being formed with evry possible fuzzy set */
        foreach $a \in \mathcal{A}$ **do**
          $loc_s \leftarrow loc_s \cup \{\langle s \cdot a \cdot it_{i+1} \rangle\}$
    $\mathcal{S}_{\mathcal{FG}} \leftarrow \mathcal{S}_{\mathcal{FG}} \cup \{loc_s\}$
end

IV. Experimental Results

A. DataSet

As said above, we consider the dataset provided online [WKZ+05]. The number of microarrays amounts to 286 and 17,816 genes are considered, which intensity ranges from -7.54 to 8.57 (meaning that the maximum difference between two gene expressions is 16.11).

The dataset aims at discovering classification tools for dicrimating between two types of cancer: agressive and non agressive). We thus took the genes for which the difference of expression is significative between these two classes by using the SAM (Significant Analysis MicroArray) method, usually used by biologists, which uses the FDR and q-value method presented in [Sto02]. 555 genes were extracted by this process.

On top of this data, we consider the following trapezoidal fuzzy sets that were provided by the biologist:

- *almost equal* (genes belong then to the same itemset) when the difference ranges from 0 to $ln(2)$
- *lightly over-expressed* (hereafter lightly) by means of a fuzzy set of kernel $[ln(2), 3 * ln(2)]$ and support [3*ln(2),6*ln(2)]
- *averagely over-expressed* (hereafter avg) by means of a fuzzy set of kernel $[6 * ln(2), 7 * ln(2)]$ and support $[3 * ln(2), 9 * ln(2)]$
- *very over-expressed* (hereafter very) by means of a fuzzy set of kernel $[9 * ln(2), \infty[$ and support $[7 * ln(2), \infty[$

B. Some Extracted Rules

We report here some rules that have been extracted by our approach and were described as interesting by the expert. For confidential reasons, we do not provide the gene name hidden behind the number we provide.

Without any management of gaps between itemsets, the following sequence was discovered: $\langle(546)(411)(51)\rangle$. However, this sequence was not relevant as genes 546 and 411 have similar expression levels for many experiments.

By considering gaps using crisp intervals, the following pattern was discovered: $\langle(546\ 411)\_lightly\_(51)\rangle$ showing the interest of considering intervals as:

- it was then possible to consider 546 and 411 as similar,
- it was possible to describe how 51 was expressed compared to these two first items.

However, it is very difficult for biologists to define crisp partitions as the cutting is too strict and does not represent the reality. Biologists thus defined a fuzzy partition, which allowed us to discover to which extent the pattern $\langle(546\ 411)\_lightly\_(51)\rangle$ was true. This support was 3.63 instead of 5 in the crisp case. This information is of great importance to the expert.

Finally, one of the more actionable results from the experiments we led comes from the fact that the experts would like to find out patterns to discriminate between the benign and malignant cases. For this purpose, we compare the patterns found from the two subsets. In our former experiments, many patterns (e.g. $\langle(5)(41)(51)\rangle$) were the same in both experimental conditions and were thus considered as non discriminant. However, by using our approach, we found out that this pattern occurs in both conditions, but with very different fuzzy gaps. Indeed, for benign tumors, the pattern $\langle(5)avg(41)lightly(51)\rangle$ can occur, whereas the pattern $\langle(5)avg(41)very(51)\rangle$ occurs for malignant tumors. Our approach can thus help discriminating between the two forms of tumors.

V. Conclusion

In this paper, we study the problem of mining biological data for discovering relevant gene interactions in the framework of breast cancer. For this purpose, we show how Fuzzy Logic is of great interest in order to better take into account. The experimental results reported here highlight these promising results.

Further work will include the study of the properties of the fuzzy constraints introduced by the fuzzy partition in order to enhance the performances of our algorithm in terms of memory and time consumption, and further experiments on other databases, including other microarray data and clinical data from a psychological study both related to the Alzheimer disease. Moreover, we aim at studying how the computation of the support impacts the sequences found by our approach. In particular, we will compare the results obtained in this paper using a sigma-count to the results obtained by a thresholded count or a thresholded sigma-count. This comparison will be led both on the performance side (time and memory consumption) and on the semantic side by an evaluation provided by the experts.

References

[AS95] R. Agrawal and R. Srikant. Mining sequential patterns. In *ICDE'95*, pages 3–14, 1995.

[BCS+05] EM Blalock, KC Chen, AJ Stromberg, CM Norris, I Kadish, SD Kraner, NM Porter, and PW Landfield. Harnessing the power of gene microarrays for the study of brain aging and alzheimer's disease: statistical reliability and functional correlation. *Ageing Res. Rev.*, 4(4):482–512, 2005.

[ESBD98] M. Eisen, P. Spellman, P. Brown, and D. D.Botstein. Cluster analysis and display of genome-wide expression patterns. In *National Academy of Science*, volume 85(25), pages 14863–14868, 1998.

[FLT07] C. Fiot, A. Laurent, and M. Teisseire. Extended time constraints for sequence mining. In *14th IEEE International Symposium on Temporal Representation and Reasoning (TIME'07)*, 2007.

[FLT09] C. Fiot, A. Laurent, and Maguelonne Teisseire. Softening the blow of frequent sequence analysis: Soft constraints and temporal accuracy. *International Journal of Web Engineering and Technology (IJWET)*, 2009.

[HDG05] F. Hoerndli, DC David, and J Götz. Functional genomics meets neurodegenerative disorders. part ii: Application and data integration. *Progress Neurobiol.*, 76:169–188, 2005.

[JG05] X.R. Jiang and L. Gruenwald. Microarray gene expression data association rules mining based on bsc-tree and fis-tree. *Data Knowl. Eng.*, 53(1):3–29, 2005.

[JKZ73] A. Jones, A. Kaufmann, and HJ Zimmermann. *Fuzzy Sets: Theory and Applications*. Masson, 1973.

[KMC00] M. Kathleen Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. In *Journal of Computational Biology*, volume 7, pages 819–837, 2000.

[MO04] SC Madeira and AL Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[Org09] World Health Organisation. Cancer. 2009.

[PBB08] RG Pensa, J Besson, and JF Boulicaut. Constrained co-clustering of gene expression data. In *SIAM International Conference on Data Mining*, 2008.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proc. of the Fifth Int. Conference on Extending Database Technology (EDBT)*, pages 3–17, 1996.

[SBT09] P. Salle, S. Bringay, and M. Teisseire. DEMON: DEcouverte de MOtifs squentiels pour les puces ADN. In *Proc. conf. EGC*, pages 459–460, 2009.

[Sto02] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:479–498, 2002.

[TTC01] V. Tusher, R. Tibshirani, and C. Chu. Significance analysis of microarrays applied to the ionizing data analysis radiation response. In *Natl. Acad. Sci.*, volume 98, pages 5116–5121, 2001.

[WKZ+05] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.