



Department of Computer Science,
University of Bari, IT

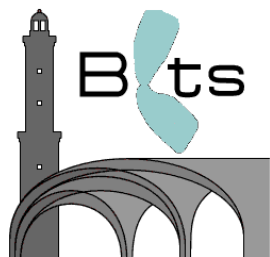


Institute for Biomedical Technologies
CNR - Bari, IT

Discovering Relational Association Rules for the Characterization of UTR cis-regulatory modules

Eliana Salvemini

Department of Computer Science University of Bari



esalvemini@di.uniba.it

domenica.delia@ba.itb.cnr.it

Research Goal

Structural characterization of translation cis-regulatory modules

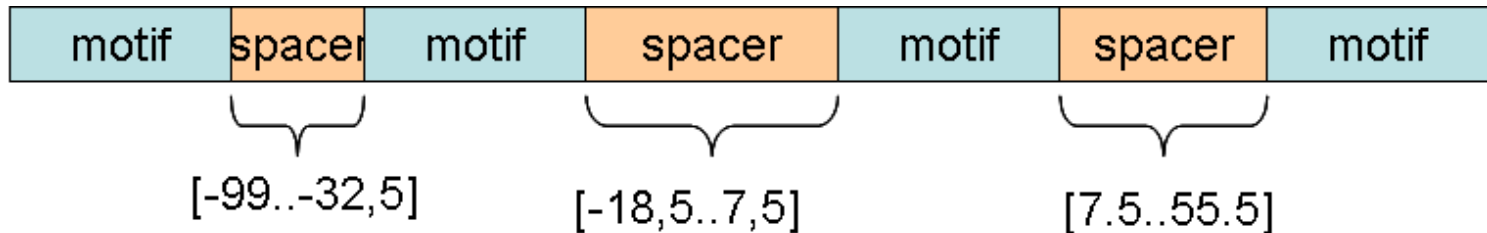
We address this biological problem by applying **data mining techniques**

Idea: discover **frequent combinations** of regulatory motifs (named **patterns**), since their significant co-occurrences could reveal important functional relationships

The data mining approach

Our approach allows to discover **spaced patterns**

- composed of two or more motifs of arbitrary length
- interleaved with spacers whose **lengths** can vary in ranges of values not defined a priori

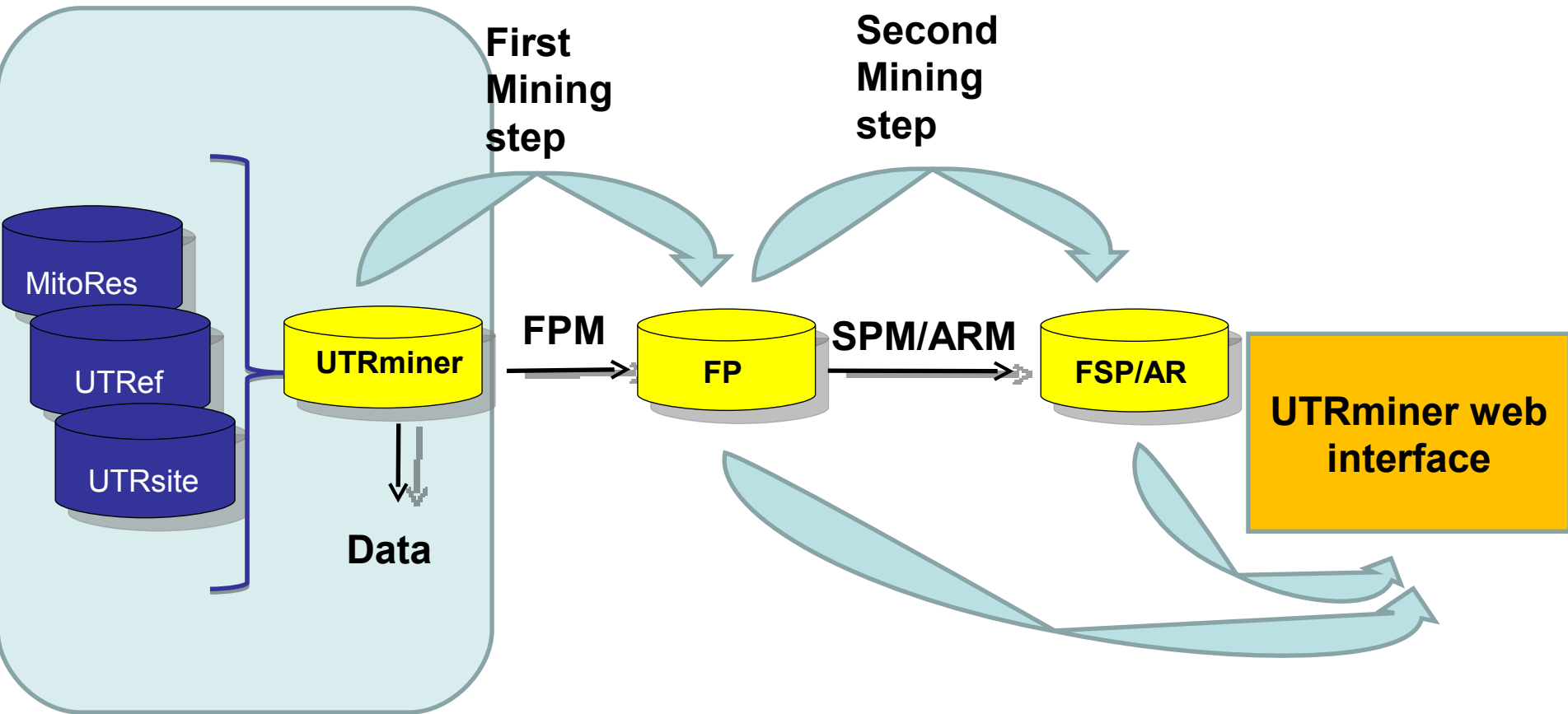


The data mining approach

A two-stepped data mining procedure:

1. mine **frequent patterns (FP)**, that is, frequent **sets** of different motifs which co-occur along the UTR sequences (their spatial displacement is not considered)
2. mine **frequent sequential patterns (FSP)**, that is, frequent **sequences** of spaced motifs, which hopefully correspond to **cis-regulatory modules**

The approach



First mining step

INPUT: a view on UTRminer which associates UTR **sequences** with their contained **motifs** and their **length**, **starting** and **ending position** in the biological sequences

- **Candidate patterns** are **sets of different motifs**
- The **support** of a candidate pattern is the number of UTRs sequences in which all motifs of the candidate co-occur
- Search starts from the smallest candidates (sets with a single motif) and proceeds towards larger sets
- A candidate pattern (set of motifs) is **frequent** (**infrequent**) if its support is higher (lower) than a minimum threshold (**minsup**)
- The set of motifs which are frequent at the i -th level are considered to generate candidate sets of motifs at the $(i+1)$ -th level

OUTPUT: a collection of **frequent patterns** (FP)

First mining step results



Contact Links

Home > 3'UTR > mt-3'UTR Human Patterns

Main Menu

* Home

UTRminer

- * Databases
- * Frequent Patterns
- * Frequent Sequential Patterns

Frequent Patterns

- * 5'UTR
- * 3'UTR
- ... mt-3'UTR Total Patterns
- ... mt-3'UTR Human Patterns



Frequent Patterns | mt-3'UTR Human Patterns

Pattern level: 1

- Pattern level: 1
- SECIS1 - [4]
 - Mos-PRE - [26]
 - ADH_DRE - [15]
 - GY-BOX - [45]
 - K-BOX - [80]
 - SXL_BS - [60]
 - CPE - [15]
 - BRD-BOX - [60]
 - UNR-bs - [50]
 - PAS - [483]
 - IRES - [148]
 - uORF - [645]

Pattern level: 2

Pattern level: 3

Pattern level: 4

Pattern level: 5

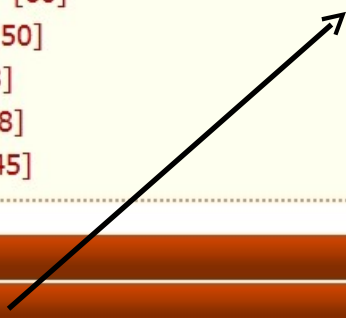
Frequent Patterns | mt-3'UTR Human Patterns

Pattern level: 1

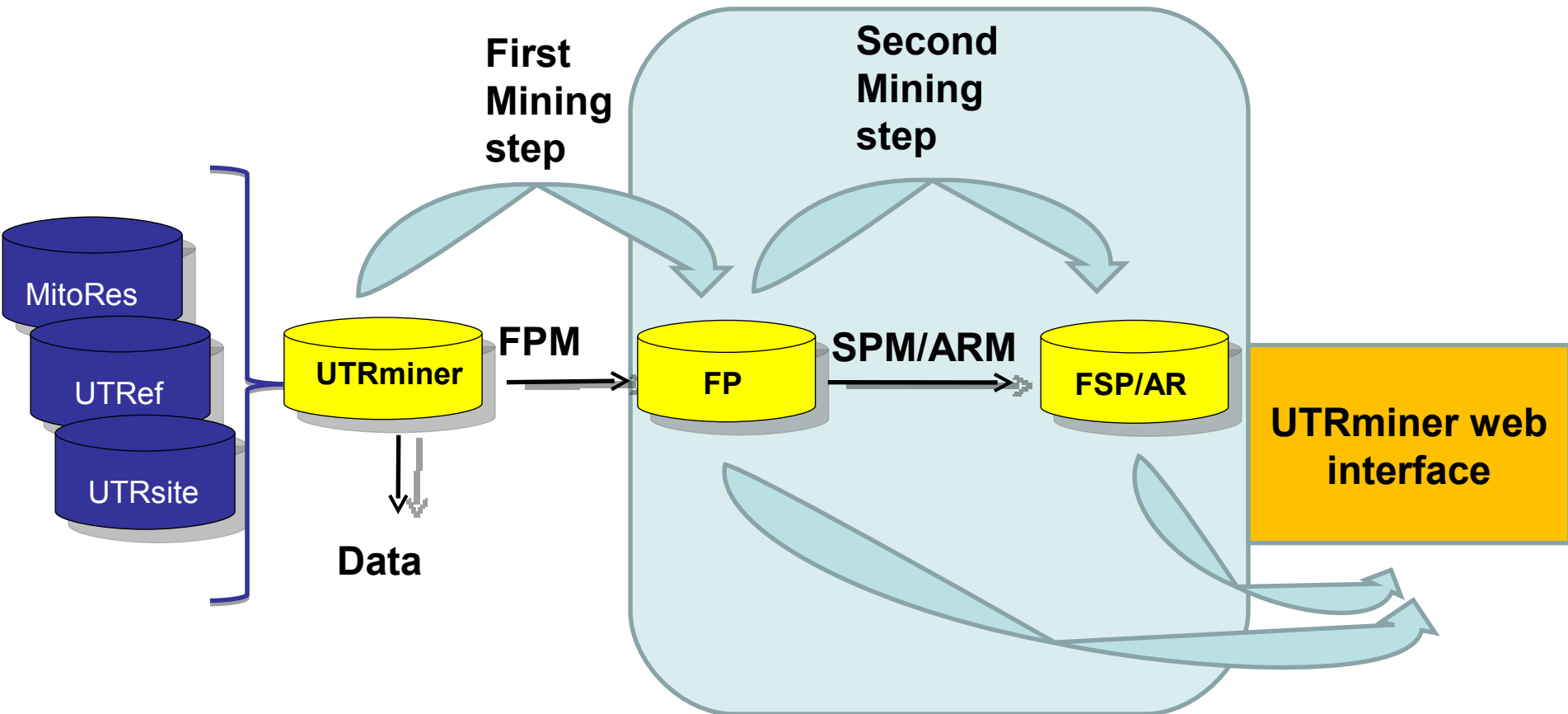
Pattern level: 2

Pattern level: 3

- Pattern level: 3
- INIT: 43 | BRD-BOX,IRES,Mos-PRE - [4]
 - INIT: 44 | BRD-BOX,K-BOX,SXL_BS - [4]
 - INIT: 45 | BRD-BOX,IRES,K-BOX - [4]
 - INIT: 46 | ADH_DRE,PAS,SXL_BS - [4]
 - INIT: 47 | CPE,SXL_BS,uORF - [4]
 - INIT: 48 | K-BOX,Mos-PRE,uORF - [4]
 - INIT: 49 | GY-BOX,IRES,PAS - [4]
 - INIT: 50 | GY-BOX,PAS,UNR-bs - [4]
 - INIT: 51 | GY-BOX,PAS,SXL_BS - [4]
 - INIT: 52 | ADH_DRE,SXL_BS,uORF - [5]
 - INIT: 53 | GY-BOX,IRES,uORF - [4]
 - INIT: 54 | Mos-PRE,SXL_BS,uORF - [5]



Second mining step



Preparing data for the second step

- For every pair of two **consecutive motifs** p_1 and p_2 the length of the **spacer in-between** is computed as the difference between the *endingPosition* (last nucleotide) of p_1 and the *startingPosition* (first nucleotide) of p_2

Example:

$$\left. \begin{array}{l} p_1: \langle p_1, 100, 200 \rangle \\ p_2: \langle p_2, 250, 300 \rangle \end{array} \right\} \rightarrow \langle p_1, p_2 \rangle = \langle p_1, 50, p_2 \rangle$$

- The length of a spacer between two motifs is a negative or positive integer depending on whether motifs overlap or not
- An UTR is modelled as a **sequence of motifs with spacers in-between**

Second mining step

- GOAL: mine **frequent sequential patterns (FSP) of motifs** also by taking the **spacer** between motifs into account
- Algorithms for FSPs can work only on discrete variables
- PROBLEM: information on spacers' length is numeric (integer)
- IDEA: discretizing **spacers' lengths**
 - partitioning the range of values into a small number of intervals (or bins), and then
 - convert spacer lengths by mapping them into their corresponding interval
- ALGORITHM: **equal frequency discretization** numerical values are approximately uniformly distributed among non-overlapping intervals of different width
- EXPERIMENTS: performed at 6, 9 and 12 bins

Discretization

Example:

- $\langle A, 30, B, 1000, C, -200, D \rangle$, sequence of spaced motifs,
- the length of spacers is discretized into three bins:
 - $[-300, -1] \rightarrow \text{NEG_DISTANCE}$
 - $[0, 210] \rightarrow \text{SHORT_DISTANCE}$
 - $[211, 1100] \rightarrow \text{LONG_DISTANCE}$
- the original sequence is transformed into the following one:
 $\langle A, \text{SHORT_DISTANCE}, B, \text{LONG_DISTANCE}, C, \text{NEG_DISTANCE}, D \rangle$
- **Frequent sequential patterns** are mined on these transformed data
- They are represented as sequences

$$\langle M_1, S_1, M_2, S_2, \dots, S_n, M_n \rangle$$

where

- M_i denotes a motif
- S_i denotes an interval returned by the discretization procedure

Second mining step: GSP

To discover FSPs two algorithms have been considered

1. GSP (Agrawal & Srikant, 1995)

- available in WEKA
- discovered patterns are not strictly sequences
A B C D → AB, AC, AD, ABC, ACD, BC, BD, BCD, CD
are all valid patterns
- In a previous work we tested GSP on nuclear transcripts targeting mitochondria from 10 different species of Metazoa (1944 5'UTR and 1952 3'UTR sequences)

Results GSP

- H-dataset: INIT 88 – FP: **PAS, IRES, uORF**
- 111 sequences

	Support 20	Support 30
Bin 6	a) uORF [-99..-18.5] IRES [-99..-18.5] PAS (47) b) uORF, [73.5..438], uORF, [41.5..73.5], uORF (27) c) uORF, [-18.5..7.5], uORF, [73.5..438], uORF (26) d) uORF, [41.5..73.5], uORF, [20.5..41.5], uORF (26) e) uORF [7.5..20.5] uORF [41.5..73.5] uORF (29)	uORF, [-99..-18.5], IRES, [-99..-18.5] PAS support (47)
Bin 9	uORF, [-99..-25.5], IRES, [-25.5..0.5], PAS support(34)	uORF, [-99..-25.5], IRES, [-25.5..0.5], PAS support (34)
Bin 12	uORF, [-99..-30.5], IRES, [-30.5..-18.5], PAS support (34)	uORF, [-99..-30.5], IRES, [-30.5..-18.5], PAS (support:34)

GSP: Issues

GSP discovers frequent sequential patterns but

- **many of them are useless** because they do not present the canonical structure

$\langle M_1, S_1, M_2, S_2, \dots, S_n, M_n \rangle$

- some FSPs do not begin and end with a motif
- motifs are not interleaved with spacers
- The discovery of FSPs is very sensitive to the discretization process

higher number of bins → $\left\{ \begin{array}{l} \text{FSPs are more specific} \\ \text{BUT} \\ \text{their support is lower} \end{array} \right.$

Second mining step: SPADA

- SPADA [Lisi & Malerba, 2004] discovers **spatial association rules** (AR)
- At first it discovers spatial patterns and then generates spatial association rules from them
- A **spatial pattern** P is a conjunction of predicates, at least one of which is a **spatial relation**
- The **support** of a spatial pattern P estimates the probability of observing P
- A **spatial association rule** $Q \rightarrow R$ is obtained from a **spatial** pattern $P=Q \wedge R$
- The **confidence** of an association rule estimates the conditional probability $P(R | Q)$
- In our application, if R represents the last motif in a sequence then the confidence is useful to make predictions on the basis of the first part of the sequence

SPADA

- The basic element of a pattern is an atomic formula (or **atom**), that is, a predicate symbol applied to some terms (variables or constants)

Example:

uORF, distance1, IRES...

utr(A),part_of(A,B), is_a(B,uorf), distance1(B,C), C\=B, is_a(C,ires)...

- SPADA performs different phases to generate AR:
 1. **Candidate generation:** Generate **candidate patterns** with k atoms from frequent patterns with $(k-1)$ atoms
 2. **Candidate evaluation:** Generate **frequent patterns** from candidate patterns with k atoms *until no more frequent patterns are found*
 3. **AR generation:** Generate **association rules** from frequent patterns

SPADA: advantages

- SPADA can exploit a domain theory expressed as Prolog programs
- We exploit this characteristic to define admissible merging of bins produced by the discretization process
- In particular, we indicate to merge n bins $[A_1, B_1], \dots, [A_n, B_n]$ iff:
 - They are consecutive, i.e., $B_i = A_{i+1}$
 - The resulting interval $[A_1, B_n]$ has a length $B_n - A_1$ which is less than a fixed number of nucleotides
- In this way SPADA can mine rules formed both by the **original bins** and by the **merged ones**
- SPADA is less sensitive to the initial discretization respect to GSP
- In SPADA it is possible to specify several constraints which prevent the generation of useless patterns, such as those generated by GSP

$\langle M_1, S_1, M_2, S_2, \dots, S_n, M_n \rangle$

SPADA: issues

The output of SPADA presents some difficulties of reading because of **the heavy redundancy of similar rules due to the merging of bins:**

→ three filters are applied to SPADA output

Filters on the AR

- Filter1: more specific rule that is the rule with the smaller bin

M1	S1	M2	S2	M3	S3	M4	Supp	Conf
uORF	[3.5.. 29.5]	uORF	[-99.. -18.5]	IRES →	[-99.. -18.5]	PAS	32.43	100
uORF	[3.5.. 29.5]	uORF	[-99.. -18.5]	IRES →	[-99.. 3.5]	PAS	32.43	100

- Filter2: the rule with greater support

M1	S1	M2	S2	M3	S3	M4	Supp	Conf
uORF	[3.5.. 29.5]	uORF	[-99..-18.5]	IRES	[-99..-18.5] →	PAS	32.43	100
uORF	[3.5.. 29.5]	uORF	[-99.. 3.5]	IRES	[-99..-18.5] →	PAS	35.13	100

- Filter3: the rule with greater confidence

M1	S1	M2	S2	M3	S3	M4	Supp	Conf
uORF	[3.5.. 29.5]	uORF	[-99..-18.5]	IRES	[-99..-18.5] →	PAS	32.43	92
uORF	[3.5.. 29.5]	uORF	[-99.. 3.5]	IRES	[-99..-18.5] →	PAS	35.13	100

Results of SPADA: init 88, 12 bin e support 30

M1	S1	M2	S2	M3	S3	M4	Supp	
uORF	[-99..-18.5]	IRES	[-99..-18.5]	PAS			47	
uORF	[-18.5..55.5]	IRES	[-99..-18.5]	PAS			37	
uORF	[-99 ..-30.5]	IRES	[-30.5 .. -18.5]	PAS			34	
uORF	[3.5..72.5]	uORF	[-99..-18.5]	IRES	[-99..-18.5]	PAS	28	
uORF	[7.5..72.5]	uORF	[-18.5..55.5]	uORF	[3.5..72.5]	uORF	49	
uORF	[-18.5..55.5]	uORF	[29.5..111.5]	uORF			64	
uORF	[20.5..55.5]	uORF	[7.5..55.5]	uORF	[7.5..72.5]	uORF	31	
uORF	[29.5..111.5]	uORF	[-18.5..55.5]	uORF	[3.5..72.5]	uORF	49	

Results of SPADA: init 88, 12 bin e support 30

M1	S1	M2	S2	M3	S3	M4	Supp	Conf
uORF	[-99..-18.5]	IRES	[-99..-18.5] →	PAS			47	100
uORF	[-18.5..55.5]	IRES	[-99..-18.5] →	PAS			37	94,87
uORF	[-99 ..-30.5]	IRES	[-30.5 .. -18.5] →	PAS			34	100
uORF	[3.5..72.5]	uORF	[-99..-18.5]	IRES	[-99..-18.5] →	PAS	28	100
uORF	[7.5..72.5]	uORF	[-18.5..55.5]	uORF	[3.5..72.5] →	uORF	49	100
uORF	[-18.5..55.5]	uORF	[29.5..111.5] →	uORF			64	100
uORF	[20.5..55.5]	uORF	[7.5..55.5]	uORF	[7.5..72.5]	uORF	31	81,57
uORF	[29.5..111.5]	uORF	[-18.5..55.5]	uORF	[3.5..72.5] →	uORF	49	100

Conclusions

- Patterns mined by SPADA can also be mined by GSP but only if the minsup is lowered. This means losing information about the significance of a pattern, because it is less supported.
- SPADA gives a further piece of information, the **confidence**, which helps to **predict the presence of a motif**, given the motifs which precede it in the sequence.
- The patterns mined by GSP are filtered because many of them don't have any sense (they aren't spaced motifs). All patterns mined by SPADA have sense, although they **must be filtered** because of their similarity.

Conclusions

- SPADA mines **classes of equivalence** of spaced sequences of motifs, each of them containing all sequences of motifs which vary not for structure but for spacer dimension.
- The **filters** serve to choose more representative patterns of each class of equivalence.
- SPADA is able to mine patterns which are **trains of motifs**, while GSP isn't (unless by significantly lowering the minsup), which means that SPADA offers major possibilities to detect sequences of spaced motifs given the same conditions.

Acknowledgements

Antonio Turi
Corrado Loglisci
Donato Malerba

Department of Computer Science,
University of Bari, IT

Giorgio Grillo
Domenica D'Elia

Institute for Biomedical Technologies
CNR, Bari, IT

UTRminer: <http://utrminer.ba.itb.cnr.it/>

Many thanks for your attention!



Institute for Biomedical Technologies
CNR - Bari, IT



Department of Computer Science,
University of Bari, IT