Exploiting a Biclustering algorithm in ORFeome analysis

Vera Afreixo
Departamento de Matemática
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
vafreixo@mat.ua.pt

Adelaide V. Freitas
Departamento de Matemática
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
adelaide@mat.ua.pt

M. Pinheiro
IEETA
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
monsanto@ieeta.pt

José L. Oliveira
IEETA
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
jlo@ua.pt

G. Moura
Departamento de Biologia
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
gmoura@bio.ua.pt

Manuel A. S. Santos
Departamento de Biologia
Universidade de Aveiro
Campus de Santiago,
3810-193 Aveiro, Portugal
msantos@bio.ua.pt

ABSTRACT

Biclustering algorithms have emerged as an important method for analyzing large scale expression data of microarrays technology. In this paper we present a modification of the existing Iterative Signature Algorithm (ISA). The proposed algorithm represents an extension of the criterium suggested by [3] for the identification of homogenous biclusters which presents a performance more stable than the ISA.

We have implemented four different bicluster algorithms (ISA, one side ISA, ISA-median, one side ISA-median) in a public domain software named Anaconda that is being used for comparative codon context analysis of all the available open reading frames in a genome (the ORFeome). Applying our algorithms in the *Saccharomyces cerevisiae* codon context map, and taking the most frequents and biggest biclusters, we identify patterns not detected before when hierarchical clustering and ISA were used. For others species we also explore some bicluster results in the analysis of codon context.

1. INTRODUCTION

The increasing number of sequenced genomes and the large amount of complex data emerging from DNA microarrays technologies, have created new challenges in several scientific domains, namely statistics and computational sciences. An important challenge, associated with biological data sets is the definition of rules or criteria that lead to the identification of interpretable patterns and homogeneous groups.

The potential of clustering methods to reveal biologically meaningful patterns was initially considered in [2], who applied hierarchical clustering to the identification of functional groups of genes. However, standard clustering techniques have shown some limitations. For example, in microarray data sets, these methods do not provide overlapping between clusters and so, they are not adequate for biological systems, where the same gene may be involved in multiple processes and therefore belong to multiple clusters.

The Iterative Signature Algorithm (ISA) is a biclustering algorithm proposed by [5, 4] for detecting groups in the microarray data matrix, considering simultaneousely the two dimension of the matrix and allowing to get overlapping groups. The ISA can be applied to other numerical data matrices since its biclustering criterium is defined in terms of averages. In fact, this algorithm was designed to find sub-matrices in the data matrix whose observations in each row and each column fails, on average, out of some interval (or rule). However, the current version of the ISA, publicly available at BicAT in http://www.tik.ee.ethz.ch/sop/bicat [1], sometimes forms unexpected groups, which are not interpretable. The criterion which defines the biclustering procedure is based on the behaviour of averages and, it is well-known that, the average is a central measure strongly influenced by errors and outliers that can be appear in biological data sets. Bearing that fact [3] proposes a modification of the ISA defining a biclustering rule in term of the medians in spite of averages. Also it is observed that the ISA-median has, in many aspects, a more stable performance than the ISA:

- the ISA-median has a symmetric feature (one can transpose the data matrix and the results are similar), the same does not apply to ISA;
- the ISA presents higher probability of finding biclusters even when there is no statistical significance;
- the ISA-median presents less quantity and bigger biclusters than the ISA and it detects more stable biclusters than the ISA;
- in general, the number of different biclusters detected

VLDB ‘07, September 23-28, 2007, Vienna, Austria.

by the ISA-median are independent of an initial parameter which is not true for the ISA;

- in the ISA, both threshold parameters have a stronger impact on the performance.

We have created a software application named Anaconda to help discovering the important features of gene primary structure, at a genomic scale. In the context of multivariate analysis, we have included in Anaconda the following biclustering algorithms: ISA-median, the ISA and one side modifications. For each specie with available ORFeome, Anaconda analyzes the bias and the association between consecutive genetic symbols (nucleotide, codon or amino acid), by building an $n \times m$ contingency tables with two variables, doing residual analysis and multivariate analysis (see [6, 7] for more details).

Recently, one goal of the Anaconda software is to identify biclusters in codon context. This will allow us to detect patterns of codon pairs which decoding by the ribosome is highly problematic. In this situation we observed than the ISA yields a high number of groups (submatrices) with one or two rows and columns which do no lead to any interpretation of the biclusters. That fact shows the influence that one observation in the context codon maps can have and hence preventing the detection of patterns or homogeneous groups in our data. Using ISA-median it does not happen.

This paper is organized as follows. In the next section we will describe the one side ISA-median biclustering algorithm in detail. In Section 3 we will present some results of biclustering on the ORFeome of *Saccharomyces cerevisiae* and we will discuss the threshold parameters of the ISA-median for different species. In the last section we briefly discuss some conclusions and future work.

2. MODIFICATIONS ON ISA BICLUSTERING ALGORITHM

Given a general numerical data matrix $\mathbf{X_{n\times m}}$. A bicluster of $\mathbf{X_{n\times m}}$ is a submatrix $\mathbf{B} = \mathbf{X_{IJ}}$ of $\mathbf{X_{n\times m}}$ with values correlated or homogeneous according to a certain criterion.

Considering the statistical meaning of clusters obtained by the ISA, we can say that the clustering criterium is based on the behavior of the sample mean assuming that the observations of the matrix $\mathbf{X}$ represents one random sample from a Gaussian population. Since the average is a central measure strongly influenced by errors and outliers, ISA can form clusters due to the interference of single values in rows and columns. However, ISA's criterium of biclustering can be modified to overcome these limitations, taking the statistical behavior of the median instead the average. Thus, in [3] we developed a modified biclustering algorithm, ISA-median.

However, such as was initialy proposed, the ISA-median fails when we intent to find homogenous groups of observations that present an unexpected behavior in one direction (one side) but not two (see figure 1). Having this in mind, we propose a new algorithm, a brief modification of the ISA-median – one side ISA-median.

In the sequel above we summarize in detail the steps of the ISA-median, considering three criterion one side (lower and higher) and two side (module).

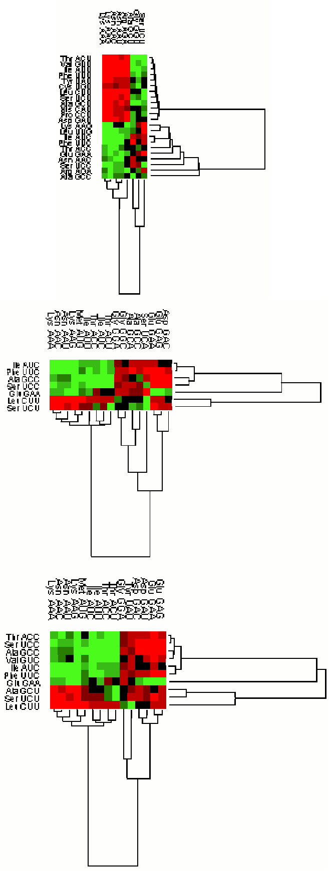

Figure 1: ***Bicluster of*** *Saccharomyces cerevisiae* ***map, obtained using ISA-median algorithm and hierarchical clustering.***

One side and two side ISA-median

Input:

- $\mathbf{X_{n\times m}}$: $n \times m$ matrix of the observations
- $C = \{C_j,\ j = 1, \cdots, m\}$ - set of m treatments of the factor *Column*;
- $R = \{R_i,\ i = 1, \cdots, n\}$ - set of n treatments of the factor *Row*;
- $R^{(0)}$ - an initial set of $n_0 \leq n$ treatments of the factor *Row*;
- T_R - threshold for the factor *Row*;
- T_C - threshold for the factor *Column*;
- $\widehat{f(\overline{S}_L)} = \frac{1}{\sqrt{2\pi}\widehat{\sigma}} \exp \quad -\frac{1}{2} \quad \frac{\overline{S}_L - \widehat{\mu}}{\widehat{\sigma}} \quad ^2$, $L \in \{R,\ C\}$, $\widehat{\mu} = \frac{\sum_{ij} x_{ij}}{n \times m}$ and $\widehat{\sigma} = \sqrt{\frac{\sum_{ij}(x_{ij} - \widehat{\mu})^2}{n \times m - 1}}$.

FIRST STAGE

Step 1: Initialize $k = 0$.

Step 2: Obtain the submatrix of $\mathbf{X}$ for the selected rows $R^{(k)}$, $\mathbf{X}_{R^{(k)}}$.

Step 3: Compute medians by columns, $S_{C_j} = med\{X_{ij} : R_i \in R^{(k)}\}$.

Step 4: Calculate the average of the medians by column, $\overline{S}_C = \dfrac{\sum_j S_{C_j}}{m}$.

Step 5: Obtain the subset $C^{(k)}$ of C which treatments C_j fall out a pattern pre-defined in terms of a threshold T_C and in the following way:

- **Module**

  $$C^{(k)} = \quad C_j \in C : |S_{C_j} - \overline{S}_C| > T_C \sigma_C$$

- **One side**
  - **Lower**

    $$C^{(k)} = \quad C_j \in C : S_{C_j} - \overline{S}_C < T_C \sigma_C$$

  - **Higher**

    $$C^{(k)} = \quad C_j \in C : S_{C_j} - \overline{S}_C > T_C \sigma_C$$

where $\sigma_C = \dfrac{1}{2\sqrt{\#\ R^{(k)}}\widehat{f(\overline{S}_C)}}$

SECOND STAGE

Step 6: Obtain the submatrix of $\mathbf{X_{n\times m}}$ for the selected columns $C^{(k)}$, $\mathbf{X}_{C^{(k)}}$.

Step 7: Compute medians by rows, $S_{R_i} = med\{X_{ij} : C_j \in C^{(k)}\}$.

Step 8: Calculate the average of the medians by rows, $\overline{S}_R = \dfrac{\sum_i S_{R_i}}{n}$.

Step 9: Obtain the subset $R^{(k+1)}$ of L which treatments R_i fall out a pattern pre-defined in terms of a threshold T_R and in the following way:

- **Module**

  $$R^{(k+1)} = \quad R_i \in R : |S_{R_i} - \overline{S}_R| > T_R \sigma_R$$

- **One side**
  - **Lower**

    $$R^{(k+1)} = \quad R_i \in R : S_{R_i} - \overline{S}_R < T_R \sigma_R$$

  - **Higher**

    $$R^{(k+1)} = \quad R_i \in R : S_{R_i} - \overline{S}_R > T_R \sigma_R$$

where $\sigma_R = \dfrac{1}{2\sqrt{\#\ C^{(k)}}\widehat{f(\overline{S}_R)}}$.

Step 10: If $R^{(k+1)} \neq R^{(k)}$ then make k equal to $k+1$ and repeat Steps 2 to 9 else stop.

Output: Bicluster $\mathbf{X_{R^{(k)}C^{(k)}}} = [x_{ij}]_{\substack{i \in R^{(k)} \\ j \in C^{(k)}}}$.

ISA-median, ISA and one side version could be seen on the hypothesis test context. For ISA-median the null hypothesis is: $\mu = \overline{S}_A$, with $A \in \{C,\ R\}$ (see algorithm step 5 and 9) and μ represents the population median. The thresholds T_C and T_R (see algorithm input) are the critical value(s) of all hypothesis test from the algorithm. The value of the test statistic, in a sample, is compared to determine whether, or not, the null hypothesis is rejected. The ISA-median is already prepared for the two-sided test or one-sided test (step 5 and step 9).

3. RESULTS

In order to identify patterns of the effect of codon context on mRNA decoding accuracy, in species with completely sequenced genome, we used the Anaconda software to built codon-pair context maps. Then, we applied ISA-median to find biclusters in each map.

We applied ISA-median and one side ISA-median to the codon-pair context maps for different species and for different values of the thresholds $T_C \in [0, 4]$ and $T_C \in [0, 4]$ (see input algorithm). We have observed that, along all the analyzed species, the thresholds that leads to the most frequent bicluster is not always the same and so it depends, as [6] shown, on species-specific codon context fingerprints.

For *Saccharomyces cerevisiae*, the map clearly shows that each codon has a set of preferred 3'-codon neighbors and rejects a set of other codons (see figure 2).

Applying hierarchical clustering allows the identification of some patterns. The main goal of this study was to identify patterns which cannot be detected with hierarchical clustering. For instance, *Saccharomyces cerevisiae* map has an important feature of gene primary structure. For example, in the figure 3 there is a relationship formed by the nucleotid in the last position of first codon and first nucleotid of the second codon (see [6]).

However, the hability of the clustering approach to detect groups is limited. Furthermore, that difficulty is even higher

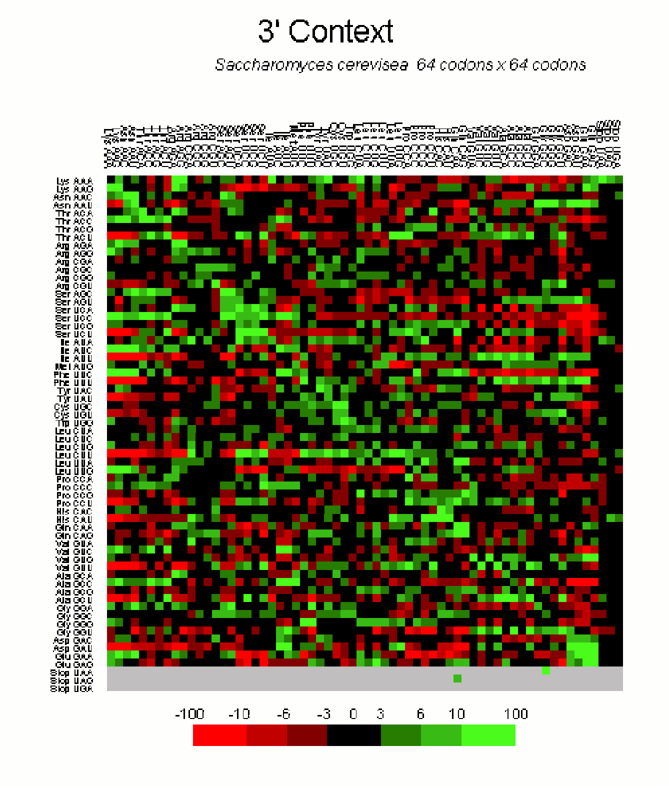

Figure 2: ***Saccharomyces cerevisiae*** **map.**

when we increase the dimensions of the data matrix. Otherwise we are interested in finding sub-matrices (groups) whit rows and columns show a similar extremal behavior.

On one hand in the ISA median, lower or higher versions, the algorithm extracts lower or higher values, according to the respective rule (see algorithm). These biclusters form homogenous groups with low values (red) or high values (green), respectively. On the other hand we use the module version (see algorithm step 5 and 9) to extract higher or lower values according to the presented rule and biclusters with high extreme value (green and red).

Herein we applied ISA-median to analyze patterns into the codon-pair context map for *Saccharomyces cerevisiae* (real data matrix), whose values are adjusted residual Pearson values (see [7]). We took into account more frequents biclusters with more then three rows and columns. In the figures 5 and 4 are some biclusters for *Saccharomyces cerevisiae* map.

The most frequent biclusters present the following characteristics:

- the nucleotide U-uracil in the third position of the first codon and the nucleotide A-adenine in the first position of the second is a highlighted bad context (similar result was observed in [6]), see figure 4 first bicluster;
- when the second codon begins with the sequence GA, i.e., XXX-GAX, it detected some bad contexts (see the last three biclusters in figure 4);
- the pair XXU-GXX presents a good strong context in

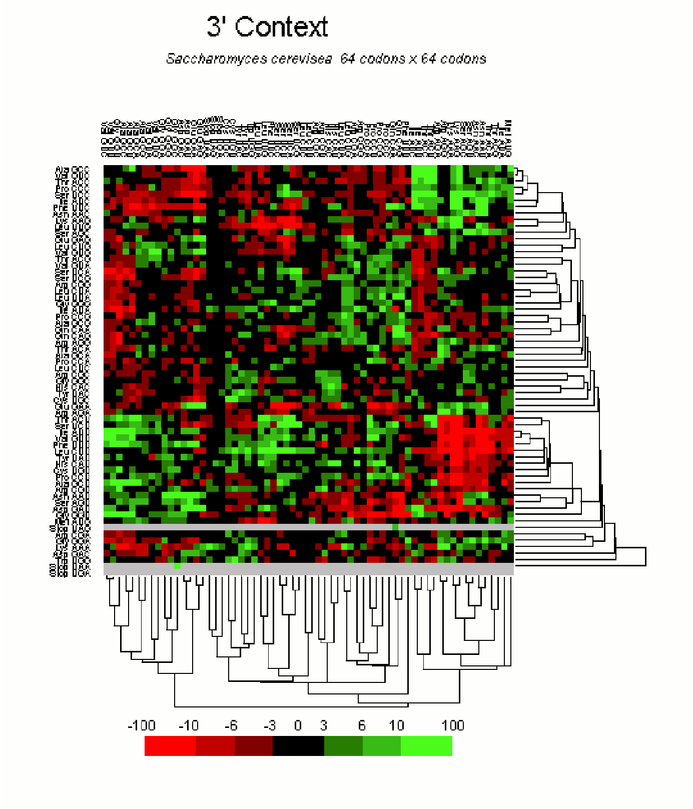

Figure 3: ***Hierarchical cluster*** *Saccharomyces cerevisiae* ***map.***

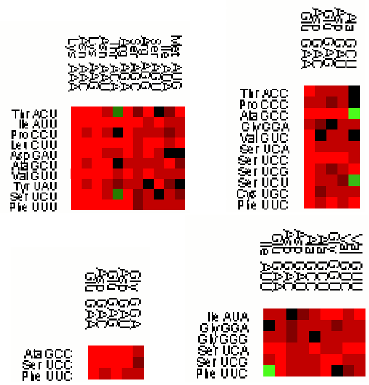

Figure 4: ***Bicluster of residual matrix of*** *Saccharomyces cerevisiae*. ***Lower.***

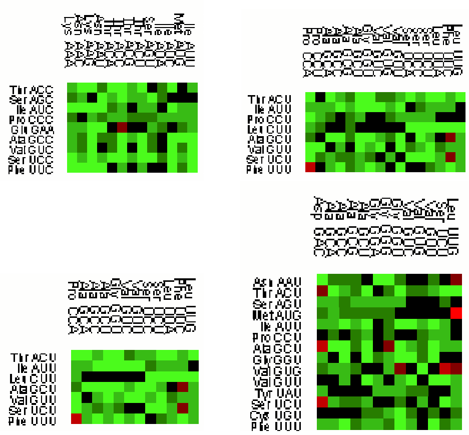

Figure 5: *Bicluster of residual matrix of* *Saccharomyces cerevisiae*. *Higher.*

many biclusters (similar result was observed in [6]), see figure 5 first bicluster;

- the XXC-AXX context (the context defined by C in the last position of the first codon and by A in the first position of the second codon) is a good strong context in many obtained biclusters, see the last three biclusters in figure 5.

We also applied ISA-median to other ORFeomes and for different threshold values and we obtained valid biclusters when we consider threshold values ranging from 1.5 to 2.5. Note that the choice of the thresholds to get the more frequent bicluster is dependent upon the genome map.

For comparing the performance of the discussed algorithms, each one was applied one hundred times to the *Saccharomyces cerevisiae* map. Table 1, presents, for some relevant thresholds, the number of different biclusters (n_m) formed by the algorithm and the absolute frequency of the most frequent bicluster (f_{max}). We observed that the ISA algorithms yields a higher number of groups and the most frequent bicluster as a small frequency (shown in italic). This does not happen with ISA-median algorithms. We could see in the table 1 that we have a smaller number of different biclusters and the most frequent bicluster has a higher frequency (shown in bold).

4. CONCLUSIONS AND FUTURE WORK

We have presented a modification of the ISA biclustering algorithm, the one side ISA-median, that extends previous work [3].

We have applied both ISA-median and one side ISA-median for *Saccharomyces cerevisiae* detecting important feature of gene primary structure that modulates mRNA decoding accuracy and that were not found before.

Despite these promising results, further work is needed. We intent to develop a methodology that can measure the significance of the results (biclustering validation) and to apply ISA-median for analyzing gene expression microarray data.

Acknowledgement

This research was supported by *Fundação para a Ciência e a Tecnologia* (Portugal) through PTDC/MAT/72974/2006 and *Unidade de Investigação Matemática e Aplicações* of University of Aveiro.

5. REFERENCES

[1] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*.

[2] M.B. Eisen, P. T. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*.

[3] A. V. Freitas, M. Pinheiro, V. Afreixo, J. Duarte, J. L. Oliveira, G. Moura, and M. Santos. A median-based iterative signature algorithm. *In Proceedings of the Statistics for Data Mining, Learning and Knowledge Extraction (accepted)*.

[4] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*.

[5] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, and Y. Ziv. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*

[6] Gabriela Moura, Miguel Pinheiro, Raquel M. Silva, Isabel M. Miranda, Vera M. A. Afreixo, GD Gaspar Dias, Adelaide Freitas, Jos Lus Oliveira, and Manuel Santos. Comparative context analysis of codon pairs on an orfeome scale. *Genome Biology*, 6(3):R28(14), 2005.

[7] M. Pinheiro, V. Afreixo, G. Moura, A. Freitas, M. A. Santos, and J. L. Oliveira. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine*, 45:163–168, 2006.

Table 1: Algorithms performance comparison.

		ISA(Module)	ISA(Lower)	ISA(Higher)	ISA-median(Module)	ISA-median(Lower)	ISA-median(Higher)
T_R	T_C	$n_m(f_{max})$	$n_m(f_{max})$	$n_m(f_{max})$	$n_m(f_{max})$	$n_m(f_{max})$	$n_m(f_{max})$
1,5	1,5	0 (0)	***22 (9)***	***42 (10)***	0 (0)	**16 (25)**	**13 (27)**
	1,8	7 (10)	***18 (8)***	***36 (12)***	1 (1)	**11 (10)**	**11 (30)**
	2	9 (13)	6 (2)	***34 (5)***	4 (8)	**7 (35)**	**10 (31)**
	2,3	***13 (6)***	6 (4)	***28 (6)***	**5 (31)**	**10 (14)**	**9 (15)**
	2,5	***14 (4)***	1 (3)	***22 (6)***	0 (0)	**5 (10)**	**6 (21)**
1,8	1,5	5 (8)	***16 (7)***	***21 (5)***	**3 (10)**	**5 (16)**	**9 (13)**
	1,8	9 (6)	9 (2)	***21 (3)***	2 (9)	**3 (27)**	**8 (13)**
	2	***14 (6)***	4 (4)	***22 (3)***	**1 (14)**	**4 (10)**	**7 (16)**
	2,3	***14 (5)***	2 (3)	***24 (4)***	**3 (20)**	**4 (17)**	**6 (19)**
	2,5	***13 (4)***	1 (1)	***15 (4)***	**2 (30)**	4 (9)	**6 (19)**
2	1,5	5 (5)	8 (11)	***19 (2)***	**1 (21)**	4 (2)	**7 (11)**
	1,8	***10 (3)***	5 (4)	***14 (2)***	**5 (29)**	**5 (14)**	**3 (12)**
	2	***14 (3)***	3 (1)	***18 (2)***	3 (3)	**4 (16)**	**4 (15)**
	2,3	7 (5)	1 (1)	***13 (5)***	2 (7)	**3 (11)**	6 (7)
	2,5	9 (4)	0 (0)	***17 (3)***	**3 (20)**	3 (4)	4 (6)
2,3	1,5	1 (1)	5 (2)	***11 (1)***	3 (3)	**4 (10)**	6 (4)
	1,8	7 (2)	1 (2)	9 (1)	**6 (23)**	6 (7)	3 (4)
	2	6 (3)	0 (0)	8 (2)	1 (1)	**4 (16)**	2 (3)
	2,3	8 (2)	0 (0)	6 (1)	2 (1)	**3 (12)**	4 (4)
	2,5	8 (3)	1 (1)	3 (1)	**5 (15)**	1 (1)	3 (5)
2,5	1,5	2 (1)	2 (1)	5 (1)	0 (0)	**5 (14)**	3 (5)
	1,8	1 (1)	0 (0)	2 (1)	**3 (19)**	**3 (15)**	3 (3)
	2	2 (1)	0 (0)	4 (2)	2 (2)	**1 (12)**	2 (5)
	2,3	4 (1)	0 (0)	6 (1)	**5 (15)**	**2 (15)**	1 (8)
	2,5	***10 (2)***	0 (0)	5 (2)	4 (6)	0 (0)	3 (7)