*ECAI 2012*
*Luc De Raedt et al. (Eds.)*

*doi:10.3233/978-1-61499-098-7-654*

# Hierarchical and Overlapping Co-Clustering of mRNA:miRNA Interactions

**Gianvito Pio**[1] and **Michelangelo Ceci**[2] and **Corrado Loglisci**[3] and **Domenica D'Elia**[4] and **Donato Malerba**[5]

**Abstract.** microRNAs (miRNAs) are an important class of regulatory factors controlling gene expressions at post-transcriptional level. Studies on interactions between different miRNAs and their target genes are of utmost importance to understand the role of miRNAs in the control of biological processes. This paper contributes to these studies by proposing a method for the extraction of co-clusters of miRNAs and messenger RNAs (mRNAs). Different from several already available co-clustering algorithms, our approach efficiently extracts a set of possibly overlapping, exhaustive and hierarchically organized co-clusters. The algorithm is well-suited for the task at hand since: *i)* mRNAs and miRNAs can be involved in different regulatory networks that may or may not be co-active under some conditions, *ii)* exhaustive co-clusters guarantee that possible co-regulations are not lost, *iii)* hierarchical browsing of co-clusters facilitates biologists in the interpretation of results. Results on synthetic and on real human miRNA:mRNA data show the effectiveness of the approach.

## 1 Introduction

microRNAs (miRNAs) are small ribonucleic acid (RNA) molecules that can be found in most of eukaryotic cells. They are post-transcriptional regulators that bind to complementary sequences on target messenger RNA (mRNAs). In the last decade, miRNAs were recognized as a distinct class of biologic regulators with conserved functions. In particular, research has revealed multiple roles in negative regulation (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation) [6, 19]. By affecting protein production of genes, miRNAs are likely to be involved in most biologic processes and control many metabolic pathways [26].

The study of the possible bonds between miRNAs with complementary sequences on target mRNAs has been recognized as an interesting biological research problem that is worth to be investigated. Indeed, miRNA expression profiles can provide valuable clues for investigating the properties of miRNAs, such as tissue specificity and differential expression in cancer/normal cells [18]. On the other hand, different mRNAs that are bound by the same miRNAs may share unknown functional properties. In this context, the application of co-clustering techniques seems to be a natural choice in order to identify co-clusters of miRNAs and mRNAs. In fact, as recognized in [5], the task of co-clustering well matches one of the major problems in computational biology: discovering regulatory modules that control gene transcription in biological model systems. As a consequence, several papers in the literature apply co-clustering algorithms in the biological domain [3, 28, 21, 4, 7]. However, they work on gene expression data and not on predictions of miRNA : mRNA interactions.

In order to properly work on miRNA:mRNA interactions, three important issues have to be considered: *i)* mRNAs and miRNAs can be involved in different regulatory networks that may or may not be co-active under all conditions [4], ignoring this aspect would lead to the identification of incomplete regulatory networks; *ii)* in order to avoid losing possible co-regulations, each miRNA and mRNA should belong to at least one co-cluster (exhaustiveness); *iii)* miRNAs:mRNAs prediction datasets are inherently large and the application of classical co-clustering techniques may result in a high number of extracted co-clusters, thus negatively affecting the interpretability of results. In order to face these three issues, it is necessary to extract overlapping and exhaustive co-clusters which are hierarchically organized (hierarchy helps biologists in the analysis of the results).

There are a few papers in the literature that extract overlapping co-clusters from gene expression data. In one of the pioneering works on this topic [4], a greedy heuristic search is performed to generate arbitrarily positioned, overlapping co-clusters, based on a homogeneity constraint. However, co-clustering is based on iterative insertions and deletions of genes and conditions asymmetrically (i.e. insertions and deletions of conditions depend on insertions and deletions of genes). Moreover, as pointed out in [7], this iterative algorithm is expensive, since it identifies individual co-clusters sequentially rather than all at once. The algorithm also causes random perturbations to the data, while masking discovered co-clusters, which reduces the co-clustering quality. In [21] genes and conditions are represented according to a binary matrix which is recursively divided into two smaller (possibly overlapping) submatrices, after a rearrangement of columns/rows. This means that this approach follows an expensive search strategy [1] that is impractical for large datasets. In [1] the authors defined a co-cluster as an order-preserving submatrix (OPSM). According to this definition, a co-cluster is a group of rows whose values induce a linear order across a subset of the columns. A submatrix is order-preserving if there is a permutation of its columns under which the sequence of values in every row is strictly increasing. As in [4] rows and columns are not interchangeable. Moreover, OPSM does not support hierarchical co-clusters and is designed to identify only a single co-cluster for each execution. In [3], the authors propose to extract overlapping and hierarchical co-clusters. However, co-clustering is non-deterministic, the hierarchy is built on a single dimension and overlapping is supported on the other dimension.

In [7], the authors propose an efficient meta algorithm (called ROCC) to co-cluster gene expression data. This algorithm works in a bottom-up fashion and merges co-clusters in order to obtain a hierar-

[1] University of Bari, Italy, email: gianvito.pio@uniba.it
[2] University of Bari, Italy, email: ceci@di.uniba.it
[3] University of Bari, Italy, email: loglisci@di.uniba.it
[4] Institute for Biomedical Technologies (Consiglio Nazionale delle Ricerche), Italy, email: domenica.delia@ba.itb.cnr.it
[5] University of Bari, Italy, email: malerba@di.uniba.it

chy of co-clusters. Merging is performed by identifying the "closest" co-clusters at each iteration. However, merging works on "relationships" (or edges) rather than objects (e.g. genes and conditions) and is based on a simple distance function between co-clusters. An additional problem is that extracted co-clusters are not exhaustive. Although this is motivated by the necessity of removing noise objects, it contrasts with one of the issues raised from the application at hand.

Although similar to ours, methods specifically designed to work with gene expression data have the specific goal of grouping together rows (columns) with similar (both strong and weak) expressions. This is different from our purposes, that is to group together miRNAs and mRNAs that show (only) reliable interactions.

According to this consideration, we propose an algorithm, called HOCCLUS (Hierarchical Overlapping Co-CLUStering), for efficient discovering of overlapping, exhaustive and hierarchically organized co-clusters according to a bottom-up strategy. Our algorithm does not focus on "relationships" in the merging process (as ROCC does) but on objects (units of analysis) directly, with the opportunity of dealing with unbalanced datasets (i.e. objects of different type participate with significantly different cardinalities in the interactions). Moreover, we use the concept of *separability* of co-clusters together with that of distance. This allows us to have co-clusters defined according to both density and distance-based criteria.

## 2 The proposed method

Co-clustering is strictly related to bipartite graph partitioning. A bipartite graph is an undirected graph where nodes are partitioned into $V_r$ and $V_c$[6] such that no edge connects the nodes in the same partition. Formally, a bipartite graph $G$ is defined as $G = (V_r \cup V_c, E)$, where $E \subseteq V_r \times V_c$, and can be represented by an adjacency matrix $A^{n\times m}$, where $n = |V_r|$, $m = |V_c|$ and $[A]_{ij}$ is the weight of the undirected edge $e_{ij} \in E$ that connects the node $i \in V_r$ to the node $j \in V_c$. Without loss of generality, we impose that $[A]_{ij} \in [0, 1]$.

Intuitively, starting from a non-overlapping co-clustering, which can be obtained by running one of the methods already available in the literature (e.g., [8, 16, 27]), our method consists of an iterative process in which, at each iteration, two phases are performed, that is, overlap identification and merging. In the former, some objects (miRNAs or mRNAs) belonging to a co-cluster can be added to another co-cluster. In the latter, co-clusters are merged when some heuristic criteria are satisfied. It is noteworthy that at each iteration several pairs of co-clusters can be merged. Moreover, at each iteration, depending on whether merging is performed, an additional level of the hierarchy may or may not be added. This iterative process stops when neither overlaps nor merges are performed in the last iteration.

Formally, the problem we intend to solve is defined as follows:
*Given:* a bipartite graph $G = (V_r \cup V_c, E)$, the corresponding adjacency matrix $A^{n\times m}$, a co-clustering quality function $q : \mathbb{C} \times [0, 1]^{n\times m} \to \mathbb{R}$ (where $\mathbb{C}$ is the set of possible co-clusters) and a quality threshold $\alpha$ for $q(\cdot, \cdot)$.
*Find:* a set of co-clusters $L_j$ for each level $j = 1, \ldots, k$ such that:
***a)*** for each set $L_j, j = 2, \ldots, k$ we have that $\forall\ C' \in L_j\ \exists\ C'' \in L_{j-1}$ such that $C'' \subseteq C'$ (hierarchical organization);
***b)*** co-clusters at the same level can share objects in $V_r$ and in $V_c$ (overlapping);
***c)*** for each object $o$ in $V_r \cup V_c$, for each level $j = 1, \ldots, k$, $\exists C' \in L_j$ for which $o \in C_{i,j}$ (exhaustiveness);

[6] Subscripts $r$ and $c$ stand for row and column, respectively. Here rows refer to mRNAs and columns refer to miRNAs (interchangeable, in HOCCLUS).

**Algorithm 1** Hierarchical and overlapping co-clustering.
**Input:** the matrix $A^{n\times m}$; the function $q(\cdot, \cdot)$

$L_1 =< C_i >_{i=1..l_1} \leftarrow non\_overlapping\_coclustering(A)$;
$k \leftarrow 1$;
**repeat**
  $< numOverlaps, L'_k > \leftarrow overlapping(L_k, A)$;
  $< numMerges, L''_k > \leftarrow merging(L'_k, A, q(\cdot, \cdot))$;
  **if** $numMerges > 0$ **then**
    $k \leftarrow k + 1$; $L_k \leftarrow L''_{k-1}$;
  **else**
    $L_k \leftarrow L'_k$;
  **end if**
**until** $numOverlaps = 0$ **and** $numMerges = 0$
**return** $L_1, L_2 \ldots L_k$

***d)*** for each co-cluster $C' \in L_j$ obtained after merging, $q(C') \geq \alpha$ (quality constraint)[7].

It is noteworthy that $L_k$ does not necessarily contain a single co-cluster, meaning that a forest of co-clusters is actually returned. This is coherent with the task at hand, where some sets of miRNAs might be totally unrelated to some sets of mRNAs. Moreover, the quality threshold implicitly determines the number of the levels and the number of final co-clusters (which should hopefully be small to facilitate co-cluster analysis). Algorithm 1 solves the considered problem: it takes as input the adjacency matrix and the quality function and returns the hierarchy of co-clusters.

In this work we use *METIS* [16] to generate an initial non-overlapping/non-hierarchical co-clustering, which can be obtained by forcing node weights such that in the same cluster both miRNAs and mRNAs should appear. METIS requires as input the number of co-clusters $l_1$. As proposed in [9], the search for the optimal $l_1$ can be reduced from the range $[1, n + m]$ to $[1, \sqrt{n + m}]$ without losing too much in the approximation. Since the problem of choosing the right number of co-clusters is mitigated by the hierarchical approach, we set $l_1$ to the maximum value in the range, i.e. $l_1 = \sqrt{n + m}$.

**Overlap Identification**

The basic assumption behind the overlap identification is that two non-overlapping co-clusters should be (linearly or not) separable in the space. According to this assumption, we identify objects belonging to one co-cluster that can be added to another co-cluster.

In particular, given two co-clusters $C_i$ and $C_j$ (belonging to the same level in the hierarchy), $i \neq j$, we identify two optimal separating hyperplanes between $C_i$ and $C_j$ by learning an SVM model for each dimension (miRNAs and mRNAs). Since our goal is not to build a good predictive classification model, but to evaluate the separability of objects belonging to different co-clusters, the objects in $C_i$ and $C_j$ are used as both training set and testing set. Misclassified objects are those which possibly belong to both considered co-clusters. Intuitively, the separating hyperplane can be interpreted as delineating the changes of the underlying data distribution between $C_i$ and $C_j$. This is coherent with studies that exploit SVMs for clustering [2]. When learning SVMs, each row (column) object is represented as its corresponding row (column) vector of $A$. The use of SVMs as discriminative methods is motivated by their recognized peculiarity in dealing with sparse data [14], that is a common situation in a miRNAs:mRNAs adjacency matrix. In this way we obtain an overlapping co-clustering, where the common objects are those objects that cannot be correctly classified by the separating hyperplane (Figure 1).

[7] At this stage we do not impose any additional condition on the quality function. However, as it will be clarified, it is the co-cluster compactness.

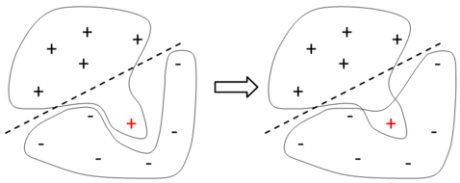

**Figure 1.** Overlapping between two clusters along one dimension. The red object (misclassified) is added to the other cluster.

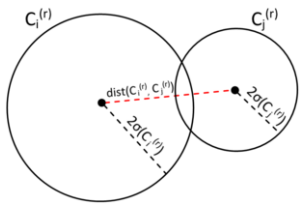


**Figure 2.** An example of the objects distribution of the row dimension of the co-clusters $C_i$ and $C_j$. In this case, $C_i$ and $C_j$ are candidates for merging.

Note that SVMs have to be constructed on each pair of co-clusters for each level. In order to obtain a result which is independent of the order in which pairs of co-clusters are analyzed, the misclassified objects are added at the end of the overlap identification process.

In Algorithm 1, $overlapping(L_k, A)$ is in charge of identifying possible overlaps. It returns the number of objects that have been added to co-clusters and the updated set of co-clusters with added objects. In our implementation, the algorithm used for learning SVMs is SMO [20] with the default kernel (linear). The usage of SMO is motivated by its linear time complexity.

### Merging

Once a set of overlapping co-clusters has been obtained, we can analyze them to evaluate if some pairs of co-clusters can be reasonably merged. A naïve approach would consider only distance or the number of common objects, neglecting the statistical distribution of the objects. Here, we assume that objects in a co-cluster are normally distributed and we consider the distance between pairs of co-clusters in order to merge those for which a defined percentage of (possibly unknown) objects can statistically be in common.

Formally, two co-clusters $C_i$, $C_j$ are candidates for merging if:

$$dist(C_i^{(r)}, C_j^{(r)}) - 2\sigma(C_i^{(r)}) - 2\sigma(C_j^{(r)}) \leq 0 \quad \textbf{or}$$
$$dist(C_i^{(c)}, C_j^{(c)}) - 2\sigma(C_i^{(c)}) - 2\sigma(C_j^{(c)}) \leq 0$$

where $C_i^{(r)}$ ($C_i^{(c)}$) is the cluster of row (column) objects belonging to the co-cluster $C_i$, $dist(x, y)$ is the euclidean distance between the centroids of the clusters $x$ and $y$ and $\sigma(x)$ is the standard deviation of the cluster $x$ (see Figure 2). Considering the factor 2 for $\sigma(x)$, we include in each sphere about 95.4% of the objects of the corresponding cluster, as a consequence of the Chebyshev's inequality.

If a pair of co-clusters $\{C_i, C_j\}$ is a candidate for merging, the quality constraint $q(C_i \cup C_j, A) > \alpha$ is evaluated. $\alpha$ allows the user to decide the minimum quality value that each co-cluster obtained after a merging step has to satisfy. Low values of $\alpha$ facilitate merging at the price of low quality co-clusters.

As quality function, we have considered the following function:

$$q(C, A) = \frac{\sum_{x \in C^{(r)}} \sum_{y \in C^{(c)}} [A]_{r(x),c(y)}}{|C^{(r)}| * |C^{(s)}|}$$

where $r : C^{(r)} \rightarrow [1, n]$ ($c : C^{(c)} \rightarrow [1, m]$) is a function that maps a row (column) object to the corresponding row (column) index of the matrix $A$. The quality function $q$ measures the intra-cluster cohesion (also known as "compactness" in classical clustering) and is computed as the normalized sum of the edge weights in $C$.

**Algorithm 2** $merging(L_j, A, q(\cdot,\cdot))$
**Input:** set of co-clusters $L_j$; the matrix $A^{n \times m}$; the function $q(\cdot,\cdot)$

```
mergeCandidates ← ∅; L ← L_j; numMerges ← 0;
for all pairs of co-clusters C_i, C_j ∈ L do
  if (dist(C_i^(r), C_j^(r)) − 2σ(C_i^(r)) − 2σ(C_j^(r)) ≤ 0 or
  dist(C_i^(c), C_j^(c)) − 2σ(C_i^(c)) − 2σ(C_j^(c)) ≤ 0) and q(C_i ∪
  C_j, A) > α then
    add < C_i, C_j, q(C_i ∪ C_j, A) > to mergeCandidates;
  end if
end for
sort mergeCandidates in descending order, w.r.t. the quality;
for all candidate cand ∈ mergeCandidates do
  if (cand.first ∈ L) and (cand.second ∈ L) then
    newCocluster ← union(cand.first, cand.second);
    remove cand.first from L; remove cand.second from L;
    add newCocluster to L;
    numMerges ← numMerges + 1;
  end if
end for
return < numMerges, L >;
```

As in the overlapping step, in order to obtain a result which is independent of the order in which pairs of co-clusters are analyzed, merging is actually performed at the end of the procedure. Obviously, a co-cluster could be a candidate for more than one merging. For example, assuming that the algorithm identifies $\{C_i, C_j\}$ and $\{C_i, C_z\}$ as two candidate pairs, we consider that with the maximum value of the quality function (see Algorithm 2).

As stated before, our overlap identification and merging procedures allow us to consider both the density of co-clusters and the distance among the objects at no additional time complexity. It can be proved[8] that our choices lead to a $max\{O(u*n*m), O(u*(n+m)^{\frac{3}{2}})\}$ worst-case time complexity ($u$ is the number of iterations).

## 3 Experiments

In this section, we evaluate the performances of the proposed algorithm (HOCCLUS) on a real dataset[9]. The dataset, which concerns the human genome, consists of a set of mRNA:miRNA pairs predictions (for a total of 13,130 mRNAs and 470 miRNAs), extracted by miRNAMap database [12]. miRNAMap collects experimentally verified miRNAs and experimentally verified miRNA target genes in human, rat and other metazoan genomes and also provides data on miRNA targets in 3'-UTR (UnTranslated Region) of genes predicted by using three algorithms (miRanda, RNAHybrid and TargetScan)[10].

The prediction strength of each mRNA:miRNA pair $(i, j)$ is described by the following three values:

- $w_{ij}^{(1)} \in \{1, 2, 3\}$ (criterion 1) is the number of algorithms which predicted the miRNA target site;
- $w_{ij}^{(2)} \in \mathbb{N}$ (criterion 2) is the number of miRNA target sites found in the same UTR region;

[8] Due to space constraints, proof is not reported in this paper.
[9] ftp://mirnamap.mbc.nctu.edu.tw/miRNAMap2/miRNA_Targets/Homo_sapiens/miRNA_targets_hsa.txt.tar.gz
[10] All the material required to replicate experiments is available at: http://www.di.uniba.it/%7Ececi/micFiles/systems/HOCCLUS/index.html

• $w_{ij}^{(3)} \in \{0,1\}$ (criterion 3) is the accessibility of the target site.

According to these criteria, the unnormalized adjacency matrix $A'$ is

$$[A']_{ij} = \gamma_1 \cdot w_{ij}^{(1)} + \gamma_2 \cdot \frac{w_{ij}^{(2)}}{\max_{\forall pairs(x,y)} w_{xy}^{(2)}} + \gamma_3 \cdot w_{ij}^{(3)}$$

where $\gamma_1, \gamma_2, \gamma_3$ are set to 1000, 10, 100, respectively. These parameters are set by the domain experts (biologists co-authors of this work) and are considered as background knowledge. The rationale is that they consider the importance of the three criteria at three different orders of magnitude, that is, criterion 1 dominates over the other two criteria while criterion 3, *ceteris paribus* on criterion 1, dominates over criterion 2. It should be noted that the proposed weighing implicitly allows us to handle noisy data (e.g. false predictions). In fact, the value of $w_{ij}^{(1)}$ can be considered as a good indicator of the confidence we have in the prediction.

Information on the position of the target site is ignored and the maximum value of the weights is considered for duplicate mRNA:miRNA prediction pairs.

The adjacency matrix $A$ is then obtained by normalizing each weight for the absolute maximum value of the matrix:

$$[A]_{ij} = [A']_{ij} / \max_{i=1,2,\ldots n; j=1,2,\ldots m} [A']_{ij}.$$

Co-clusters are evaluated on the basis of the average co-clustering compactness, which measures the strength of the intra-co-clusters connections: $p(L,A) = \frac{1}{\sum_{C_i \in L} |C_i|} \sum_{C_i \in L} |C_i|\, q(C_i, A)$, where $L$ is the set of co-clusters obtained at the last iteration and $r(x)$ and $c(y)$ are the mapping functions defined in Section 2. We notice that the average compactness is biased towards small co-clusters. This justifies the consideration of an additional evaluation measure, called average co-cluster co-regulation, which is defined as follows:

$$g(L,A) = \frac{1}{\sum_{C_i \in L} |C_i|} \cdot \sum_{C_i \in L} \frac{|C_i|}{|C_i^{(c)}|} \sum_{y \in C_i^{(c)}} s(y, A, C_i)$$

where $s(y, A, C_i)$ is the percentage of mRNAs that are bound by both the miRNA $y$ and at least another miRNA in $C_i$. Formally,

$$s(y,A,C_i) = \frac{\left|\{x \in C_i^{(r)} \,|\, [A]_{r(x),c(y)} > 0, \exists y' \in C_i^{(c)}, y' \neq y,\ [A]_{r(x),c(y')} > 0\}\right|}{\left|\{x \in C_i^{(r)} \,|\, [A]_{r(x),c(y)} > 0\}\right|}$$

The average co-cluster co-regulation is particularly useful to biologists since it allows them to identify co-regulations of mRNAs from miRNAs. Indeed, while $p(\cdot,\cdot)$ prefers small co-clusters, $g(\cdot,\cdot)$ is biased towards large ones. Our goal is to keep a good trade-off between compactness and co-regulation with a limited number of co-clusters.

Results are collected by considering the quality function defined in Section 2 and by using different values of the threshold $\alpha$.

**Quantitative evaluation**

Results reported in Table 1 show that the number of obtained levels is relatively low. Moreover, a closer analysis of the obtained hierarchies shows that they are generally balanced. Both aspects guarantee a high level of interpretability of the hierarchies.

The dataset is also analyzed with ROCC [7], that returns overlapping and hierarchical co-clusters. This analysis led to obtain a set of co-clusters that include only 33% of mRNAs and 53% of miRNAs (co-clusters are non-exhaustive) with low average compactness and co-regulation (0.010 and 0.015, respectively). The poor compactness results obtained by ROCC are motivated by the fact that ROCC extracts co-clusters with a highly balanced number of rows and columns. Although this property is desirable in principle, it does not enable ROCC to respect the original proportion among rows and columns (in the dataset we have 470 miRNAs and 13,130 mRNAs). Moreover, ROCC groups together rows (columns) with similar (both strong and weak) expressions and does not group together miRNAs and mRNAs that show (only) reliable interactions.

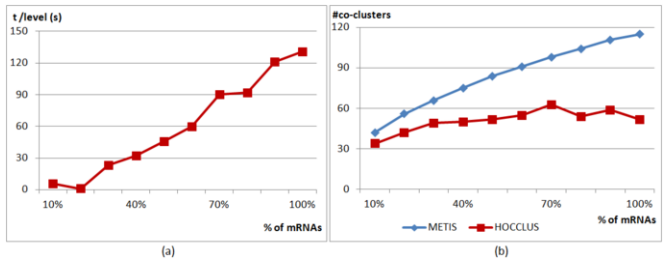


**Figure 3.** (a) Average elapsed time for each iteration, with different percentages of the number of mRNAs. (b) Number of co-clusters obtained at the last iteration with different percentages of the number of mRNAs.

**Table 1.** Results obtained by varying $\alpha$. *iter* is the number of executed iterations, *lev* is the number of hierarchy levels, *#cc* is the number of co-clusters at the last level. Comparison with $METIS$ (*lev* = 1) and ROCC is reported.

| | $\alpha$ | iter | lev | #cc | $p(\cdot,\cdot)$ | $g(\cdot,\cdot)$ | t(s)[11] |
|---|---|---|---|---|---|---|---|
| $HOCCLUS$ | 0.1 | 16 | 6 | 21 | 0.096 | 0.853 | 1711 |
| | 0.2 | 13 | 6 | 52 | 0.198 | 0.794 | 1709 |
| | 0.3 | 15 | 5 | 85 | 0.290 | 0.753 | 2767 |
| | 0.4 | 16 | 4 | 98 | 0.339 | 0.739 | 3364 |
| | 0.5 | 15 | 4 | 104 | 0.362 | 0.734 | 3322 |
| METIS | - | - | - | 115 | 0.412 | 0.746 | 10 |
| ROCC | - | - | - | 198 | 0.010 | 0.015 | 1517 |

**Table 2.** Scalability test on different portions of the dataset. The number of sampled miRNAs is kept constant (470). The first column represents the percentage of sampled mRNAs for each sampling.

| % | iter | lev | #cc | $p(\cdot,\cdot)$ | $g(\cdot,\cdot)$ | t(s)[11] | t/iter |
|---|---|---|---|---|---|---|---|
| 10% | 10 | 4 | 34 | 0.146 | 0.295 | 58 | 6 |
| 20% | 11 | 4 | 42 | 0.161 | 0.854 | 143 | 13 |
| 30% | 15 | 4 | 49 | 0.161 | 0.000 | 348 | 23 |
| 40% | 16 | 5 | 50 | 0.169 | 0.820 | 499 | 32 |
| 50% | 16 | 6 | 52 | 0.162 | 0.783 | 750 | 46 |
| 60% | 14 | 6 | 55 | 0.182 | 0.802 | 844 | 60 |
| 70% | 17 | 5 | 63 | 0.168 | 0.737 | 1503 | 90 |
| 80% | 18 | 6 | 54 | 0.175 | 0.796 | 1618 | 92 |
| 90% | 12 | 6 | 59 | 0.186 | 0.755 | 1491 | 121 |
| 100% | 13 | 6 | 52 | 0.175 | 0.794 | 1709 | 131 |

Another important aspect is that METIS reveals its ability to obtain a set of co-clusters with the best values of compactness and co-regulation. However, it cannot define a hierarchy of co-clusters, which is the actual advantage of our method. Although in Table 1 only the results obtained at the last hierarchy level are reported, the proposed algorithm returns all the identified hierarchy levels (including the first level, obtained by METIS), with a (generally) decreasing average compactness. In fact, this aspect gives the biologists the possibility to browse hierarchies of co-clusters of miRNAs and mRNAs whose higher levels consist of a relatively small number of co-clusters, each of which represents similar co-regulation roles.

Observing the overall results, in this dataset, the combination that presents the best trade-off between compactness, co-regulation and the number of co-clusters is that with $\alpha = 0.2$. As already stated, $\alpha = 0.2$ guarantees that each co-cluster obtained by merging two co-clusters has at least a compactness value of 0.2. Finally, running times are comparable with those observed for ROCC.

[11] Experiments are run on a 4 Intel CPUs @4Ghz system, 16GB of RAM.

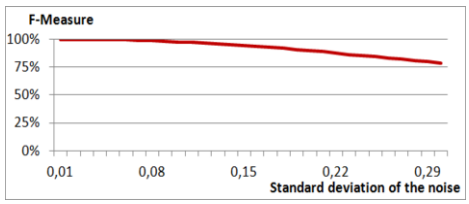


**Figure 4.** Average F-Measure (among levels) by varying the noise.

We used $\alpha$=0.2 to execute a scalability test, in order to empirically evaluate the time complexity with respect to the number of miRNAs and mRNAs. It is noteworthy that the test can be performed with any $\alpha$ value since this parameter does not directly affect time complexity. In this test, three samples obtained with a uniform sampling over the mRNAs are used and the average results are collected (see Table 2).

From the results in Table 2 and from Figure 3(a), it is possible to observe that the average running time for each iteration increases coherently with the time complexity reported in Section 2.

Finally, as it can be observed in Figure 3(b), the number of co-clusters identified at the last iteration is almost constant and does not increase considerably when the number of miRNAs increases. This is a positive aspect, which allows us to conclude that the number of co-clusters is generally independent from the number of objects, but mainly depends on objects distance and density.

**Significance of the extracted hierarchies**

In this section, we evaluate the significance of the hierarchy extracted by HOCCLUS. To this end, we generate a set of synthetic datasets, each of which includes 256 column and 16128 row objects. An initial dataset is generated by simulating a hierarchy in the underlying distribution of the data. This is obtained by imposing the strongest edge weight between objects of the 128 groups ($l_1 = \sqrt{n+m} = 128$) at the first level (for each group, we have 2 column and 126 row objects). Seven additional levels are simulated by imposing decreasing weights such that the compactness at the last level is greater than 0.2. We expect that the entire 8-levels hierarchy would be completely discovered with $\alpha = 0.2$. Starting from this initial dataset, 30 datasets are generated by adding a noise $\sim N(0, \sigma^2)$, with different values of $\sigma$. This allows us to evaluate how our method is robust to noise.

Results are measured in terms of the average F-Measure, by considering as expert's judgment the hierarchy imposed in the initial (noise-free) dataset. It is noteworthy that average F-Measure implicitly evaluates overlapping, since it is defined for multi-class classification problems. Results show (see Figure 4) that the hierarchy structure is almost correctly discovered, even in the case of very high noise ($\sigma = 0.3$ for a range of the weights of [0,1]).

**Qualitative evaluation**

In this section, the effectiveness of the algorithm in extracting biologically relevant co-clusters is presented. At this purpose, two different criteria are used: *i)* the similarity of co-clustered miRNAs on the basis of their classification in the same miRNA family or gene cluster[12]; *ii)* validated information on functional associations of co-clustered miRNAs and mRNAs. The main resources used for the analysis of co-clusters are Rfam database [11], for miRNA family classification, miRBase [17], to search complete information about miRNAs, PubMed [10], to search for literature references and DAVID [13], for functional classification of co-clustered mRNAs.

We verify whether the merging step identifies significant pairs of co-clusters to be merged. As an example[13], we now examine the co-clusters with ID "50" and "52" belonging to the level 0 of the hierarchy (identified by METIS) and the co-cluster with ID "50-52"[14] obtained by merging them during the first iteration, with $\alpha = 0.2$. We chose co-cluster 50-52 since it is one of the co-clusters with the highest compactness/co-regulation. Quantitative information on these co-clusters are shown in Table 3.

The first analysis, carried out by using miRBase, shows that, except *hsa-miR-372*, the miRNAs in the co-clusters 50 and 52 are grouped coherently with their family. As for genomic clustering, it follows more or less the same trend. Exceptions are *hsa-miR-520e*, which does not belong to any miRNAs gene cluster, and *hsa-miR-372*, which belongs to a separate cluster.

This initial analysis confirms that the algorithm used for the non-overlapping co-clustering step (METIS, in this case) provides us a good starting point for our algorithm. However, the evaluation of the co-cluster 50-52 requires a more detailed analysis. In particular, miRBase annotates for miRNAs in the co-cluster 50-52 only one useful cross-reference with [24]. In this paper, published in 2004, authors report about the results of a cDNA cloning study of miRNAs, extracted by human embryonic stem (hES) cells. They have discovered a series of new miRNAs that are expressed by human cells in the first step of embryonic development as unique and highly specialized gene set. This includes all the miRNAs grouped in the co-cluster 50 and *hsa-miR-372*, that initially belongs to the co-cluster 52.

Furthermore, in [23], the authors report that *miR-302* and *miR-372* promote human fibroblasts reprogramming to pluripotent stem cells through the targeting of multiple mRNAs . However, these findings still do not explain why *hsa-miR-372* is in the co-cluster 52, why the co-cluster 50-52, obtained by merging co-clusters 50 and 52, has so high compactness and co-regulation values and why *hsa-miR-520e*, that is not included in the miRBase gene cluster of *hsa-miR-520b*, *hsa-miR-520c* and *hsa-miR-526b** is in the co-cluster 52. By extending the search of correlations among *hsa-miR-302*, *hsa-miR-372* and *hsa-miR-520* miRNAs to specialized web resources, we find that miRDB[25], a database on miRNA target predictions and functional annotations, reports that *hsa-miR-302a, hsa-miR-302b, hsa-miR-302c, hsa-miR-302d, hsa-miR-372, hsa-miR-520b, hsa-miR-520c* and *hsa-miR-520e* share the same seed sequence.

Finally, we find confirmation of functional associations of co-clustered miRNAs of the co-cluster 50-52 in [22], in which the results of a study on the differential expression of miRNAs during the differentiation of hES cells are described. The authors demonstrated that, among all the miRNAs differentially expressed in undifferentiated hES cells, the members of the miR-302 cluster on chromosome 4 and miR-520 cluster on chromosome 19 were highly expressed. The members of these two clusters share a consensus 7-mer seed sequence and their targeted genes had overlapping functions. All these findings together fully confirm our predictions and completely satisfy the question above. By using DAVID for the functional classification of mRNAs of the co-clusters 50, 52 and 50-52, we find that some of them fall into common KEGG [15] pathways. This confirms the *ability* of the algorithm to group together mRNAs putatively involved in the same pathways, on the basis of their association with miRNAs.

[12] It has been proved that highly related miRNAs are organized as gene clusters, transcribed as polycistronic primary transcripts, and may act on the same mRNA or on different mRNAs with conserved binding sites [24].

[13] As it will be clear from the rest of the section, it is not possible to analyze several cases because of the inherent complexity of the biological analysis.

[14] The co-cluster ID is **only** an identifier. It is not used by HOCCLUS and brings with it information about its original co-clusters (after merging).

| ID | #mRNAs | $p(\cdot,\cdot)$ | $g(\cdot,\cdot)$ | miRNAs |
|---|---|---|---|---|
| 50 | 126 | 0.784 | 1.000 | hsa-miR-302a, hsa-miR-302b, hsa-miR-302c, hsa-miR-302d |
| 52 | 126 | 0.723 | 1.000 | hsa-miR-372, hsa-miR-520b, hsa-miR-520c, hsa-miR-520e, hsa-miR-526b* |
| 50-52 | 280 | 0.602 | 1.000 | hsa-miR-302a, hsa-miR-302b, hsa-miR-302c, hsa-miR-302d, hsa-miR-372, hsa-miR-520b, hsa-miR-520c, hsa-miR-520e, hsa-miR-526b* |

**Table 3.** Quantitative information about the co-clusters 50 and 52 (level 0) and the co-cluster 50-52 (level 1). Note that the number of mRNAs in 50-52 is greater than the sum of mRNAs of the co-clusters 50 and 52. This is due to the overlapping step that added some mRNAs from other co-clusters.

## 4 Conclusions

In this paper we have presented a co-clustering algorithm that faces issues coming from the study of miRNA:mRNA relationships. The proposed algorithm discovers overlapping, exhaustive and hierarchically organized co-clusters, from an initial set of non-hierarchical and non-overlapping co-clusters. The algorithm is iterative and, at each iteration, possible overlaps between co-clusters are identified and then pairs of co-clusters are merged together when some heuristic criteria are satisfied. Possible overlaps are identified through an SVM-based algorithm and merging exploits statistical properties in the data. Merging defines the hierarchical organization of co-clusters.

Experiments on human miRNAs:mRNAs data extracted from the miRNAMap database show that the proposed algorithm allows us to extract a relatively small number of co-clusters that preserve both compactness and co-regulation. A detailed biological analysis confirms that co-clusters extracted from our algorithm represent meaningful biological correlations between miRNAs and mRNAs. In particular, the comparison of extracted co-clusters with data reported in the literature (especially on known miRNA families) confirms that the algorithm may represent an effective and efficient tool for discovering unknown functional synergies among miRNAs belonging to different *families* at higher levels of the hierarchy. This would give the biologists the opportunity of discarding meaningless hypotheses whose (in-lab) validation is expensive and, on the other hand, to concentrate on the verification of potentially valid ones.

Although the proposed method has been motivated by specific needs in the biological domain, the application to artificially generated data proves that it is general enough to be used in other domains. In the future, we plan to explore this opportunity. We also plan to embed an algorithm for the automatic determination of the merging threshold $\alpha$. Finally, we intend to extend the qualitative evaluation to higher levels of the hierarchy.

**Acknowledgments**. This work is partial fulfillment of the objective of the FAR project "MBLab: Laboratorio di Bioinformatica per la Biodiversità Molecolare".

## REFERENCES

[1] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini, 'Discovering local structure in gene expression data: the order-preserving submatrix problem', in *Proc. of RECOMB '02*, pp. 49–57, (2002).

[2] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik, 'Support Vector Clustering', *Journal of Machine Learning Research*, **2**, 125–137, (2001).

[3] Jos Caldas and Samuel Kaski, 'Hierarchical generative biclustering for microrna expression analysis', in *Research in Computational Molecular Biology*, volume 6044 of *LNCS*, 65–79, (2010).

[4] Yizong Cheng and George M. Church, 'Biclustering of Expression Data', in *Proc. of ISMB'00*, pp. 93–103, (2000).

[5] Francesca Cordero, Ruggero G. Pensa, Alessia Visconti, Dino Ienco, and Marco Botta, 'Ontology-Driven Co-clustering of Gene Expression Data', in *Proc. of AI*IA 2009*, LNCS, pp. 426–435, (2009).

[6] Q Cui, Z Yu, E.O. Purisima, and E Wang, 'Principles of microRNA regulation of a human cellular signaling network', *Mol Syst Biol*, **2**, 46, (2006).

[7] Meghana Deodhar, Gunjan Gupta, Joydeep Ghosh, Hyuk Cho, and Inderjit S. Dhillon, 'A scalable framework for discovering coherent co-clusters in noisy data', in *Proc. of ICML'09*, p. 31, (2009).

[8] I. S. Dhillon, 'Co-clustering documents and words using bipartite spectral graph partitioning', in *Proc. of SIGKDD'01*, pp. 269–274, (2001).

[9] Grzegorz Góra and Arkadiusz Wojna, 'RIONA: A Classifier Combining Rule Induction and k-NN Method with Automated Selection of Optimal Neighbourhood', in *ECML'02*, LNCS, pp. 111–123, (2002).

[10] DS. Greenberg, 'National Institutes of Health moves ahead with "PubMed Central".', *Lancet*, **9183**(354), 1009, (1999).

[11] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and SR. Eddy, 'Rfam: an RNA family database.', *Nucleic Acids Res.*, **1**, 439–441, (2003).

[12] S. D. Hsu, C. H. Chu, A. P. Tsou, S. J. Chen, H. C. Chen, P. W. Hsu, Y. H. Wong, Y. H. Chen, G. H. Chen, and H. D. Huang, 'miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes', *Nucleic Acids Res*, **36**, D165–D169, (2008).

[13] DW Huang, BT Sherman, and RA Lempicki, 'Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.', *Nature Protoc.*, **1**(4), 44–57, (2009).

[14] Thorsten Joachims, 'Optimizing search engines using clickthrough data', in *Proc. of SIGKDD '02*, pp. 133–142, (2002).

[15] M Kanehisa and S. Goto, 'KEGG: Kyoto Encyclopedia of Genes and Genomes.', *Nucleic Acids Res.*, **1**(28), 27–30, (2000).

[16] George Karypis and Vipin Kumar, 'A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs', *SIAM J. Sci. Comput.*, **20**, 359–392, (1998).

[17] A. Kozomara and S. Griffiths-Jones, 'miRBase: integrating microRNA annotation and deep-sequencing data', *Nucl. Acids Res.*, **39**, D152–D157, (2011).

[18] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, 'Identification of Tissue-Specific MicroRNAs from Mouse', *Current Biology*, **12**(9), 735–739, (2002).

[19] R.F. Place, L.C. Li, D. Pookot, E.J. Noonan, and R. Dahiya, 'MicroRNA-373 induces expression of genes with complementary promoter sequences', *Proc Natl Acad Sci U S A*, **105**(5), 1608–13, (2008).

[20] John C. Platt, *Fast training of support vector machines using sequential minimal optimization*, 185–208, MIT Press, Cambridge, USA, 1999.

[21] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, 'A systematic comparison and evaluation of biclustering methods for gene expression data', *Bioinformatics*, **22**(9), 1122–1129, (2006).

[22] J Ren, P. Jin, E. Wang, FM. Marincola, and DF. Stroncek, 'MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells.', *J Transl Med.*, **7**(20), (2009).

[23] D. Subramanyam, S. Lamouille, RL. Judson, JY. Liu, N. Bucay, R. Derynck, and R. Blelloch, 'Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells.', *Nat Biotechnol.*, **29**(5), 443–448, (2011).

[24] MR Suh, Y Lee, JY Kim, SK Kim, SH Moon, JY Lee, KY Cha, HM Chung, HS Yoon, SY Moon, VN Kim, and KS Kim, 'Human embryonic stem cells express a unique set of microRNAs.', *Dev Biol.*, **2**(270), 488–498, (2004).

[25] Xiaowei Wang., 'miRDB: A microRNA target prediction and functional annotation database with a wiki interface.', *RNA.*, **14**(6), 1012–1017, (2008).

[26] R. Wilfred Bernard, Wang Wang-Xia, and Peter T. Nelson, 'Energizing miRNA research: A review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways.', *Mol Genet Metab*, (2007).

[27] Jiho Yoo and Seungjin Choi, 'Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds', *Inf. Process. Manage.*, **46**, 559–570, (2010).

[28] Sungroh Yoon, Luca Benini, and Giovanni De Micheli, 'Co-clustering: a versatile tool for data analysis in biomedical informatics.', *IEEE Trans. on inf. technology in biomedicine*, **11**(4), 493–494, (2007).