Non-coding RNA Research 6 (2021) 159–166

Contents lists available at ScienceDirect

Non-coding RNA Research

journal homepage: www.keaipublishing.com/en/journals/non-coding-rna-research

A holistic miRNA-mRNA module discovery

Ghada Shommo [a,*], Bruno Apolloni [b]

[a] Sudan University of Science and Technology, Department of Information Technology and Computer Science, Sudan
[b] Department of Computer Science, Via Comelico 39/41, 20135, Milano, Italy

ARTICLE INFO

Keywords:
miRNA-mRNA modules
Non-differentially-expressed genes
Non-predicted targets
Holistic approach
Indirect targeting
Hausdorff linkage

ABSTRACT

The regulatory role of the Micro-RNAs (miRNAs) in the messenger RNAs (mRNAs) gene expression is well understood by the biologists since some decades, even though the delving into specific aspects is in progress. In this paper we will focus on miRNA-mRNA modules, where regulation jointly occurs in miRNA-mRNA pairs. Namely, we propose a holistic procedure to identify miRNA-mRNA modules within a population of candidate pairs. Since current methods still leave open issues, we adopt the strategy of postponing any decision on the value of the module ingredients exactly at the end, i.e. at the moment of biologically exploiting the results. This diverts chains of statistical tests into sequences of specially-devised-evolving metrics on the possible solutions. This strategy is rather expensive under a computational perspective, so needing implementations on HPC. The reward stands in the discovery of new modules, possibly hosting non differentially expressed miRNAs and mRNAs and pairs containing genes that currently are considered not targeted. In the paper we implement the procedure on a Multiple Myeloma dataset publicly available on GEO platform, as a template of a cancer instance analysis, and hazard some biological issues. These results, jointly with the normal manageability of the computations, suggest that the discovery procedure may be profitably extended to a wide spectrum of diseases where miRNA-mRNA interactions play a relevant role.

1. Introduction

MicroRNAs (miRNAs) are small non-coding RNAs, found in plants, animals, and some viruses, that can cause mRNA degradation and translational inhibition, as well as mediate stimulation of gene expression. Their regulatory mechanisms in development and cellular homeostasis are still considered open issues. In particular, this paper focuses on the module discovery. According to the biological observation, multiple miRNAs may regulate one message and one miRNA may have several target genes conversely [25]. This entails bipartite miRNA-mRNA regulatory graphs, denoted *modules*, as in Fig. 1 that are associated with different *conditions* such as pathologies or histological origins [19].

The distinguishing features that are used to identify miRNA-mRNA pairs are essentially two: binding motif and expression. The former represents a rather *mechanical* feature of the ribonucleic acids, concerning small chunks of their primary structure and associated secondary structure. If they prove appropriate, essentially by complementarity, a specific miRNA may bind its mRNA partner. The appropriateness may be established algorithmically, but the effectiveness should be proven experimentally. Checked effectiveness apart, in this way we may *compute* mRNA putative target genes for any miRNA [10]. For example, miRBase has deposited 5071 miRNAs and their target genes from 58 species [21] up to date.

The miRNA and mRNA expressions constitute companion strings of values that put in relation the regulatory *activity* of the former with its effect on the mRNA in a series of patients. There are full series of experimental data that depend on the mentioned conditions [1].

1.1. Previous studies

The bipartite graph in Fig. 1 introduces the co-regulatory problem at the core of this paper as an instance of bi-clustering analysis [6]. If we consider a data matrix with row headlined by miRNAs and column headlined by mRNAs, we must identify optimal pairs of subsets of rows which exhibit similar behavior across a subset of columns, and vice versa. The main decisions in the way to get a solution concern: i) the content of the data in the matrix, and ii) the optimization procedures.

As for the former, in principle each cell of the matrix reports a distance, in a proper metric, between the crossing elements, so that the

* Corresponding author.
E-mail addresses: ghada.shummo49@gmail.com (G. Shommo), apolloni@di.unimi.it (B. Apolloni).

https://doi.org/10.1016/j.ncrna.2021.09.001
Received 26 May 2021; Received in revised form 20 September 2021; Accepted 20 September 2021
Available online 1 October 2021

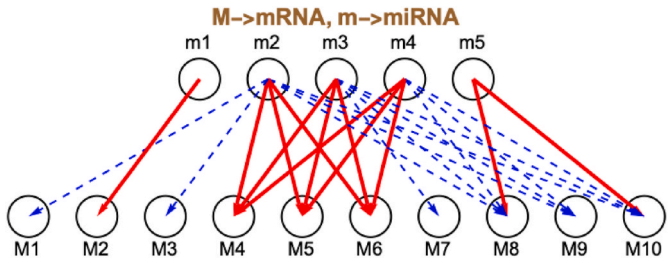

Fig. 1. Bipartite miRNA-mRNA graphs hosting *modules* aka the biclicque emphasized in red.

solution emerges from proper sub-grouping of the line and column elements. For instance, according to Bryan et Al. [6], data in the cell are the correlations between row miRNA and column mRNA, and groups are arranged so as to minimize the minimum of each subgroup, if we look for an inhibitory effect of miRNAs, under constraints on their sizes. In general, the search for the optimal grouping results in a high computational complexity problem which falls in the NP-hard class [12]. Hence, the adopted optimality criterion and related optimization algorithm play a crucial role to handle feasible procedures. For example, Bryan et Al [6]. optimize the above criterion via a BUBBLE bi-clustering strategy and the stochastic optimization method denoted as simulated annealing (SA) [30]. Liu and Tsykin [18] use binary distances, so that the cross elements come into play if they are linked by a binding motif and the correlation of their expressions is negative, discarded otherwise. Then, maximal bicliques identify modules, where a biclique is a subgraph of the original bipartite graph such that every vertex of the first partition is connected to every vertex of the second one [9]. Other authors opt for probabilistic models [20] or for soft computing techniques, such as evolutionary algorithms [17] or rough clustering [7], to *learn* approximate solutions of the optimal bi-clustering.

Finally, other authors adopt a *divide et impera* strategy, by dividing the problem in subtasks. For instance, Jayaswal et Al. [16]. identify relevant miRNA clusters and mRNA clusters first, then they look for the more significant pairs. We may see it as a three blocks procedure, where each block is a separate clustering task ending with the pruning of the candidate solutions via statistical tests of the hypothesis "no meaningful candidate", with a conventional significance level α.

1.2. Our contributions

Tough complying with the Jayaswal et Al. thread, our strategy adopts a holistic approach, as for operational aspects. In fact, we postpone pruning as latest is possible and base it on dry usability considerations. Namely:

1. unlike the common practice of focusing on only differentially expressed miRNAs and mRNAs, we assume all items available in the optioned database as candidate elements of the wanted modules.
2. in place of shrinking sets of candidate items, we maintain all of them, but progressively information enrich them with measures that denote their fitness with the module discovery goal.
3. the output of the procedure is a list of modules which is sorted according to the optimality criteria driving our path to their discovery.

The idea is that, rather than taking decisions on the basis of the partial information available at the end of the single tasks (preprocessing included), we maintain the log of the question points via preferability measures of the related options, and take decisions just at the end, when all information is available and doubts may be removed. A similar strategy is made feasible by the availability of large computational resources, so that we may manage in HPC centers large amounts of data to get the final decisions. For instance, working with the Multiple Myeloma dataset available on GEO platform,[1] we maintain alive along the procedure 296 × 7325 miRNA-mRNA pairs. As a result, we widen the scope where to find out new modules, which in turn lead to consider unprecedented pairs. Those pairs join possibly non-differentially expressed items that currently are not acknowledged as partners of a targeting. Nevertheless, early inspections of disease databases reveal the candidate relevance in regulatory phenomena.

The paper is mainly focused on computational aspects of the proposed procedure. Its organization is the following. Section 2 recalls the statistical tools used in our procedure. Section 3 illustrates the procedure. Section 4 reports its implementation and related numerical results. Finally, in Section 5 we discuss the relevance of those results from a biological perspective, hazard some specific issues and provide forewords.

2. The involved statistical tools

Identifying regulatory miRNAs and their target mRNAs is a major combinatorial challenge: in fact, a single miRNA regulates multiple mRNAs and, on the other hand, several miRNAs co-regulate a single mRNA. As mentioned in the Introduction, rather than facing directly the combinatorial problem of bi-cliquing, we prefer ordering the candidate modules according to some statistics on their components. On their basis, we: 1) look for suboptimal solutions that result computationally feasible, and 2) enable the user to bargain by himself the exhaustiveness of the solutions' set with their effectiveness.

2.1. Metrics

The metric at the basis of the above statistics is a hybrid one, as it is based on both binding motif and expression. Namely,

- We rely on $Y \times X$ binary matrices (map matrices) derived from databases and tools on the WEB, where Y denotes mRNAs, X denotes miRNAs and cell value is 1 if crossing row and column bind, 0 otherwise (the *mechanical* measure).
- Jointly, we rely on $Y \times T$ and $X \times T$ matrices on the WEB, where T spans the expression of the row headline with patients. These values are floating point numbers that have been properly normalized according to standard steps, for instance those available in NCBI platform.[2]

From the first matrix we derive a partitioning tree metric, where at the node k, seen as binary vectors, two rows (or two columns) fall in the same partition if the respective k-th bits coincide. The iterated application of this rule leads to a tree where on node h rows are located having the progressively involved h bits coinciding.

Then,

- *Individually*, from $\star \times T$ matrices we derive:
  - s1 an early similarity measure $s_{i,j}$ between rows $\{i, j\}$ by simply considering their distances, possibly in norm $L2$ – Euclidean distance, or in norm $L1$ – Manhattan distance.
  - s2 a similarity significance measure $\sigma_{i,j}$ between two rows $\{i, j\}$ as $1 - p_{value}$ of the linear regression of row i on row j (or vice-versa, since p_{value} is a symmetric function of the two rows).[3]
- *Jointly*, from $Y \times T$ and $X \times T$ matrices we derive a new $Y \times X$ similarity matrix Q, where cell $\{h, k\}$ reports:

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi.
[2] https://www.ncbi.nlm.nih.gov.
[3] http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis.

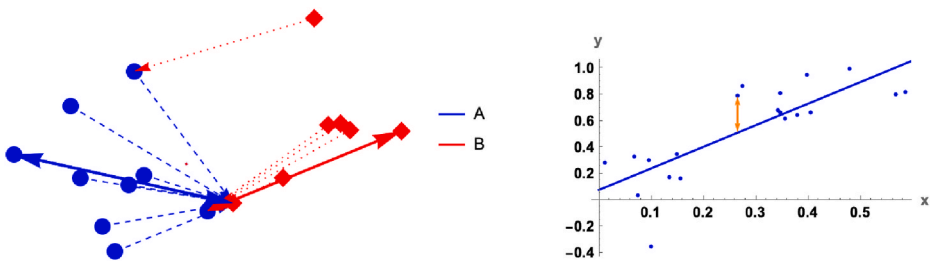

Fig. 2. A sketch of the Hausdorff distance computation in our implementation. Left picture: bullets → elements of set A, rhombuses → elements of set B; blue dashed lines → minimal distances of bullets from the set B, red dotted lines → minimal distances of rhombuses from the set A; thick double arrows Maximal distances of A from B (in blue) and of B from A (in red). Right picture: points → experimental (x,y) pair; line → their regression line; orange double arrow → the difference $\tilde{y}_i - y_i$ relative to the pair of components (x_i, y_i).

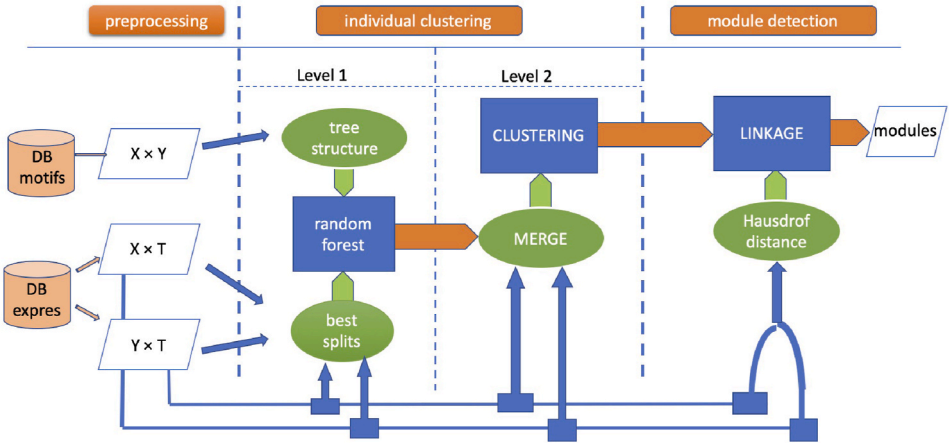

Fig. 3. The flow chart of the proposed procedure.

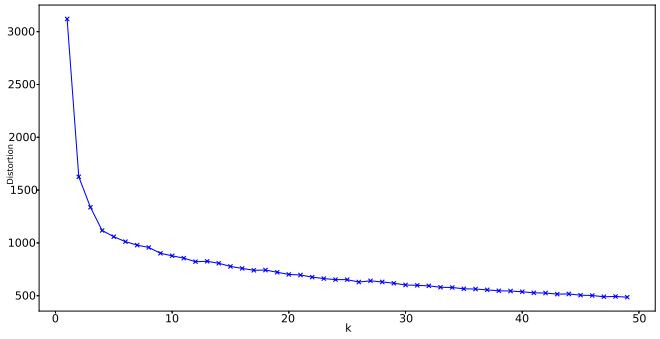

Fig. 4. The elbow graph to identify a suitable number k for clustering miRNAs.

s3 $\tilde{\sigma}_{h,k} = 1-$ the p_{value} of the regression of the mRNA headlining the h-th row of the first matrix on the miRNA headlining the k-th row of the second matrix.

Compositions of the above distance/similarity measures are at the basis of the clustering procedures used in our search path for modules.

A special mention we do to our implementation of Hausdorff distance [2, Ch. II]. By definition, the Hausdorff distance $d_H(A, B)$ is computed between two subsets A and B respectively of Y and X. In our case, the subsets are clusters of mRNAs and miRNAs, respectively, so that the distance is defined as

$$d_H(A,B) = \max\left\{ \max_{y\in A} d(y,B),\ \max_{x\in B} d(A,x) \right\} \tag{1}$$

where $d(y, B) = \min_{x\in B} d(y, x)$, $d(A, x) = \min_{y\in A} d(y, x)$, *min* and *max* are the usual *minimum* and *maximum* operators over sets, respectively, and the external maximum in (1) is carried out to take into account the non-symmetry of the two operators $\max_{y\in A}\{\min_{x\in B} d(y,x)\}$ and $\max_{x\in B}\{\min_{y\in A} d(y,x)\}$. With reference to Fig. 2-left, $d_H(A, B)$ is the length of the longest double arrow (the blue one). Moreover, differently from what represented in the above figure, in our implementation $d(y, x)$ is the Euclidean norm of the difference between y and its linearly regressed value $\tilde{y}$ over x. Namely, with reference to the set of expression values x relative to a given miRNA and the corresponding set of expression values y relative to a given mRNA, we compute the linear regression function ℓ of y over x, then we compute the regressed value $\tilde{y}_i = \ell(x_i)$, where i ranges over the joint indexes of the components of y and x . Whenever necessary, we denote this distance as d_l. With this notation, $d_l(y,x) = |y - \tilde{y}|_2$. In Fig. 2-right $\tilde{y}_i - y_i$ is represented by the double arrow in orange.

2.2. Algorithms

We implemented three clustering procedures:

Table 1
miRNA clusters (*miXX*) and mRNA clusters (*mrYY*) generated by our procedure and their sizes.

cls_name	mi0	mi1	mi2	mi3	mi4	mi5	mi6	mi7	
cls_size	4	32	23	5	58	20	23	130	
cls_name	mr0	mr1	mr2	mr3	mr6	mr7	mr9	mr10	mr11
cls_size	7	3925	1244	8	60	326	52	34	38
cls_name	mr12	mr13	mr14	mr15	mr16	mr17	mr18	mr19	
cls_size	31	1397	14	64	28	81	5	8	

Table 2
Hausdorff distances of the pairs miRNA-mRNA clusters.

	mr0	mr1	mr2	mr3	mr6	mr7	mr9	mr10	mr11	mr12	mr13	mr14	mr15	mr16	mr17	mr18	mr19
mi0	0.3346	0.3003	0.3003	0.322	0.3216	0.3093	0.322	0.3229	0.3165	0.3274	0.3052	0.3284	0.3122	0.316	0.3178	0.3373	0.3277
mi1	0.3597	0.3356	0.3346	0.3562	0.345	0.3318	0.3438	0.3562	0.3552	0.3508	0.3344	0.3427	0.3501	0.3541	0.3438	0.356	0.3541
mi2	0.1545	0.1302	0.1329	0.1527	0.1414	0.1302	0.1465	0.1371	0.1427	0.1374	0.1302	0.1532	0.1369	0.1474	0.1388	0.1572	0.1463
mi3	0.1605	0.1605	0.1605	0.1603	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1604	0.1605	0.1601	0.1603
mi4	0.3605	0.3237	0.3378	0.3637	0.346	0.3407	0.3457	0.3533	0.3434	0.3465	0.3252	0.3643	0.3444	0.3519	0.3509	0.361	0.3538
mi5	0.1605	0.1605	0.1605	0.1603	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1605	0.1604	0.1605	0.1601	0.1603
mi6	0.3277	0.3007	0.2959	0.3267	0.3125	0.3141	0.3277	0.3153	0.3143	0.3159	0.3073	0.3219	0.3221	0.3207	0.3174	0.3348	0.3193
mi7	0.3532	0.3267	0.3156	0.352	0.3401	0.3339	0.3466	0.3435	0.3499	0.3389	0.3335	0.3512	0.346	0.3404	0.3355	0.352	0.3503

C1 Hierarchical divisive clustering [23]. We use a very elementary implementation whose dendrogram results in a binary tree. Starting from a set of $q \leq 2^m$ binary strings of fixed length n, we may split it in n different ways according to the value the ith bit, for i $\in \{1, \ldots, n\}$. Iterating the procedure on each partition (son) of the split we may come to at most m partitions containing a single string. We may decide stopping the partitioning when the son meets some conditions, for instance its size is less than a threshold, and consider it a cluster. Besides these conditions, the quality of the cluster depends on the selection of the further splitting bit of a son. In absence of ancillary information the selection criterion is normally related to entropic properties of the questioned son and its prongs. In principle we should toss this criterion on each splitting bit; normally we test only on a subset of them. Random forest [27] is an ensemble of these dendrograms whose results are properly merged.

C2 Agglomerative clustering [33]. The most familiar algorithm within this family is the k-mean algorithm, where, starting from k more or less random attractors as their centers, clusters progressively grows and update their centers in a competitive way (a new point is aggregated to the closest center (the mean) and updates its value). Many variants concern the updating rule; in our method we implement the k-medoid clustering, where medoid plays the same role of mean, but refers to an actual element x of the cluster χ, the one which minimizes the sum of distances from the others. In formulas:

$$x_{\text{medoid}} = \arg\min_{y \in \chi} \sum_{i=1}^{|\chi|} d(y, x_i). \tag{2}$$

C3 Hausdorff linkage [3]. In a very essence, we use the Hausdorff distance to rank the links between miRNA and mRNA clusters in a pair.

3. The holistic procedure

Fig. 3 sketches our procedure. It consists of three phases, *pre-processing*, *individual clustering* and *module detection*, which progressively exploit the mentioned matrices $Y \times X$, $Y \times T$ and $X \times T$ derived from the WEB.

3.1. Pre-processing

For given pathology, we assess the $Y \times X$ matrix starting from the miRNA collection and companion mRNA collection available on the NCBI website and searching for targets checking which item of the latter is a target of one of the former using *mirWalk* database available online [11]. $Y \times T$ and $X \times T$ are companion matrices reporting properly normalized mRNA and miRNA expressions, respectively, of a set of patients suffering the questioned pathology at a different level.

As mentioned in the introduction, we consider all miRNAs and mRNAs at the basis of the experimental data. However, for computational reasons we may downsize their number. We do it for mRNAs by modulating the threshold on the p_value of their differential expression, where higher values than the conventional 0.05 release higher numbers of items.

3.2. Individual clustering

This phase is devoted to identify relevant clusters, individually within a mRNA dataset and a miRNA dataset. We divide this task in two: i) preparing ingredients for a good metric and ii) assembling the metric and exploiting it to identify the clusters.

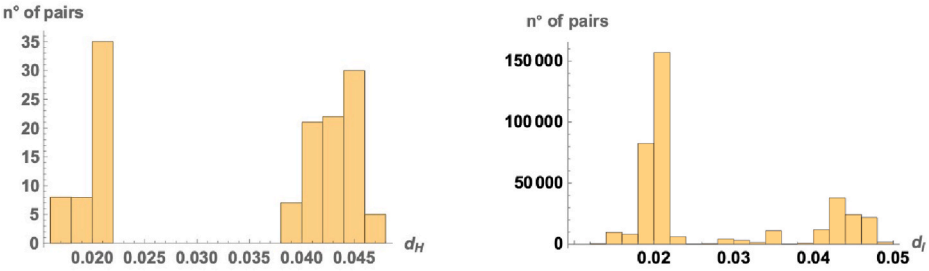

Fig. 5. Histograms of Hausdorff distances d_H and distances d_l in our case study. The former refers to all miRNA-mRNA clusters, the latter to a down sample of the miRNA-mRNA pairs. The distances have been divided by $\sqrt{n}$, where $n = 60$ is the number of the *case* patients.

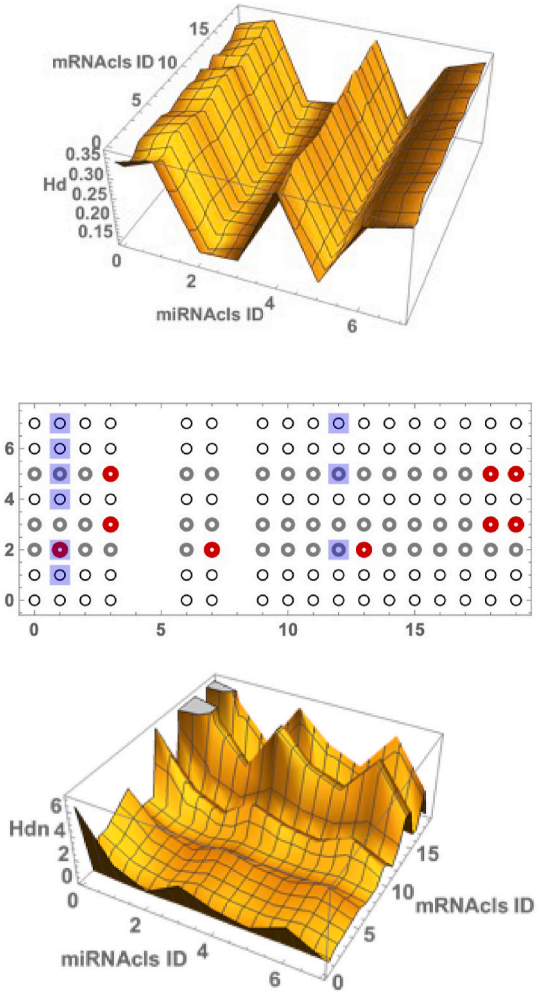

Fig. 6. Synopses of the H_d distances. Upper: the H_d landscape over the cluster pairs. Center: Marking the closest cells of Table 2: gray circles → small distance cells; red circles → smallest distance within the gray cell rows; blue squares → the 9 smallest distances according to the landscape on the bottom. Lower: same as on the upper but with reference to a *normalized* H_d.

- ***Preparing ingredients.*** Our goal is to cope with the two ways edges sketched in the bipartite graph of Fig. 1. Hence we look for Y clusters that collect mRNAs that have common targeting miRNAs (co-targeting), with high frequency, and close expression numbers (co-expression) – *analogously for X clusters.*

The co-targeting is determined by the dendrograms, let's call them *trees*, in point C1. At their completion, in a terminal node (a leaf) we find mRNAs that are targeted by a same subset of miRNAs. The splitting bit, i.e. the miRNA column in the $Y \times X$ separating mRNA rows whose cross cell in the table is labeled by 0 from those whose cell label is 1, is established at each son split by an expression homogeneity criterion. Namely, let $\mathbf{t}_i$ be the i-th row of the $Y \times T$ matrix, i.e. the expression values of the i-th mRNA. Assume that a son Q in the tree contains mRNAs $\{y_1, \ldots, y_m\}$ and its prongs P_1, P_2 after split contain mRNAs $\{y_1, \ldots, y_k\}$ and $\{y_{k+1}, \ldots, y_m\}$, respectively. We assume as non-homogeneity measure $\omega(A)$ of set A the quantity

$$\omega(A) = \sum_{i \in |A|} s^2_{i,*} \tag{3}$$

where s is the similarity measure mentioned on point s1, index * points at a dummy row representing the average of the mRNA rows in A and $|A|$ lists the indexes of those rows.[4] The homogeneity criterion sorts the candidate bits b_ks (all bits but the ones used along the branch ending with the questioned node) via the differential homogeneity measure $\Delta\omega(S, b_k)$ given by

$$\Delta\omega(Q, b_k) = \omega(Q) - \omega(P_{k_1}) - \omega(P_{k_2}) \tag{4}$$

where we further index the prongs with k to relate them to the splitting bit. Moreover, in order to limit the computational load we adopt the strategy of considering only a random subset of the candidate bits (for instance, with cardinality order of the square root of their number, according to Ref. [32]). The bit with highest $\Delta\omega$ is used to split the son Q.

At the end of the procedure, that occurs when all candidate prongs have cardinality less than a threshold τ, we remain with a set of leaves that we may handle as the results of an unsupervised clustering. Actually, Xiao and Segal [32] revisited this procedure as a multivariate regression tree [8] within the CART family [5], where $Y \times X$ rows or columns are used as independent variables and the expressions $\mathbf{t}_i$s as dependent variables. The latter does not supervise the clustering; rather, $\mathbf{t}_i$s feed an unsupervised clustering in line with [28], where the authors introduce a dummy target of the regression algorithm that results in a homogeneity measure.

Finally, the drawbacks deriving from limiting the number of candidate splitting bits is relieved by a huge repetition of the procedure generating multiple trees depending on the sampled candidate-bit-subsets. This is a usual technique denoted as *random forest*. Parameters of the obtained forest are:

1. the threshold τ to the size of the prongs
2. the size μ of the candidate subsets
3. the number ν of the trees.

- ***Final clusters*** The similarity measure through which to cluster the miRNAs and the mRNAs comes from a synthesis of the random forest results plus a further contribution from the item expressions. Namely:
  - on the one hand we merge the leaves of the trees by the formula

[4] Thus ω(A) is the sum the trace elements of the covariance matrix of those rows.

Table 3
Main statistical features of extremal modules. C miRNAs → miRNAs inside the cluster, C mRNAs → mRNAs inside the cluster; M pairs → pairs inside the module, *t* pairs → targeted pairs inside the module; DE miRNA/mRNA → differentially expressed miRNA/mRNA.

Module	{2,7}	{2,1}	{2,13}	{3,3}	{3,18}	{3,19}	{5,3}	{5,18}	{5,19}	{4,14}	{4,3}	{4,18}	{4,0}	{1,0}	{1,10}	{1,3}	{1,18}	{1,11}
No of C miRNAs	23	23	23	5	5	5	20	20	20	58	58	58	58	32	32	32	32	32
No of C mRNAs	326	3925	1397	8	5	8	8	5	8	38	8	5	7	7	34	8	5	38
No of M pairs	7498	90275	32131	40	25	40	160	100	160	2204	464	290	406	224	224	256	160	1216
Rate of *t* pairs	0.875	0.608	0.846	0.125	0.040	0.025	0.287	0.34	0.500	0.824	0.838	0.751	0.830	0.727	0.727	0.769	0.750	0.754
No miRNAs in *t* pairs	23	23	23	2	1	1	17	16	19	58	58	57	57	32	32	31	32	31
No DE miRNAs in *t* pairs	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
No DE miRNAs in *nt* pairs	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
No mRNAs in *t* pairs	326	3925	1397	4	1	1	8	5	8	38	8	5	7	7	7	8	5	38
No mRNAs in *nt* pairs	315	3925	1356	8	5	8	8	5	8	38	8	5	7	7	7	8	5	38
No DE mRNAs in *t* pairs	178	1948	737	3	1	0	4	5	5	19	4	5	3	3	3	4	5	19
No DE mRNAs in *nt* pairs	171	1948	717	4	5	5	4	5	5	19	4	5	3	3	3	4	5	19

$$\psi(x_1,x_2)=\frac{1}{\nu}\sum_{h=1}^{\nu}\delta_h(x_1,x_2) \text{ where } \delta_h(x_1,x_2) = \begin{cases} 1 & \text{if } (x_1,x_2) \text{ belong to a same leaf of the } h-\text{th tree} \\ 0 & \textit{otherwise}. \end{cases} \quad (5)$$

– on the other hand we enrich this measure with the significance of the pair. Namely, we multiply $\psi(x_1, x_2)$ by the similarity measure $\sigma_{1,2}$ defined in point s2.

In this way, we obtain for both $Y \times T$ and $X \times T$ a similarity matrix W whose cell $W_{i,j} = \psi_1(x_i, x_j)\sigma_{i,j}$. This matrix is at the basis of the k-medoid clustering recalled in Section 3. It is an unsupervised clustering algorithms that requires, to be implemented, the setting of the number c of clusters to be identified. We establish this number through a common elbow procedure.[5]

3.3. Hausdorff linkage

At the end of the second phase, we are left with individual clusters whose elements are gathered together because both are related to each other via co-expression and are intertwined by elements of the complementary side of the graph in Fig. 1. To link clusters of one side to those of the other one we may rely on distance/similarity measures, such as $\widetilde{\sigma}_{h,k}$ defined in point s3, or d_l defined on page 7, between each pair of elements of each pair of clusters. We opted for the last measure which proved to be more effective. To wrap-up d_l on a pair of clusters we use the Hausdorff distance in the way we mention in Section 2. This distance is used as the linkage between the clusters, so that the sorting of the cluster pairs according to it denotes a preference direction for analyzing candidate modules. We start from the closest pairs (according to this distance) and move ahead until this analysis provides interesting results.

As anticipated since the abstract, we do not proceed by acceptation tests, decreeing which one passes them and which one does not. Rather we assemble a set of modules that may be huge in number, since we do not waste anything. Then we provide a metric according to which we move from those that we expect to be more interesting to analyze toward the less interesting ones. The interest threshold is up to the user.

4. Implementing the holistic procedure

Implementing the procedure requires feeding the data matrices, fine tuning some parameters and formatting the results.

4.1. The datasets

We focus on the Multiple Myeloma data available on the GEO page GSE16558.[6]

Namely mRNA and miRNA expression profiles come from the collection GPL8965, series GSE16558 referring to some different levels of Myeloma pathology. This series gathers:

- mRNA expression profiles obtained by using Affymetrix Gene Chip1.0 ST, in number of 33297. The expressions have been collected on 60 patients plus 5 control patients. Data normalization is carried out through: Robust multi-array (RMA) background correction, quantile normalization, median polish algorithm [15].
- companion miRNA expression profiles obtained using TaqMan low density arrays, in number of 365. The expressions have been normalized using small-nucleolar RNAs, RNU44 and RNU48, as housekeeper, and *delta CT* method [21].

[5] https://bl.ocks.org/rpgove/0060ff3b656618e9136b.
[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16558.

From the above mRNAs we drew 7325 *loosely differentially expressed* profiles using NCBI GEO2R analysis tool [31] with a p_value = 0.1. As for miRNAs we used all profiles in principle, given their short number. However, their number reduces, because only 330 appear in the mirWalk database we used to find their targets.

From these profiles we drew a $Y \times X$ matrix having dimensions 7325 × 296, because some columns result to be empty according to the above database. We remark that in this way we scale up of a factor higher than 20 with respect to the analogous matrix restricted so the sole 0.05 differentially expressed items.

4.2. Parameters

Willing to avoid the computational complexity of an exact solution of the biclicque problem, our procedure provides approximate solutions, whose success depends on the tuning of some operational parameters. Namely, we are called to establish the following parameters.

As for level 1 clustering:

1. threshold τ. We established $\tau = 10$ to threshold the size of the leaf of the trees of the random forest. We establish this value as a compromise between the gathering power of the leaf and its significance as for number of splits, hence of miRNA-mRNA interactions.
2. numcov μ. We establish μ to be equal the square root of the length of the binary strings to be split, hence 85 for mRNAs and 17 for miRNAs. It derives from a compromise, again, between computational loads and exhaustiveness of the search for the optimal splits.
3. numtrees ν. The value $\nu = 100$ is a way of further relieving the possible drawbacks coming from an inadequate value of μ.

As for level 2 clustering:

4. numclsters k. This parameter is a typical plague of agglomerative clustering. Establishing the number k of clusters is a problem of complexity comparable with the one of the whole clustering task. We decided to adopt an elbow method [4] to override this impasse. Namely, we adopted the clustering distortion as a fitness measure of a given k selection, where distortion is measured as the sum of squared distances of samples to their closest cluster center. Then we graph the trend of this measure with k and loosely identify the *elbow* of the graph (see Fig. 4). In this way we decided to choose $k = 8$ for miRNA clusters, and k = 20 for mRNA clusters, that is around twice the elbow point, to favor a meaningful split of the huge dataset.

4.3. Numerical results

From a dry computational perspective, our results consist of:

1. A set of mRNA clusters and miRNA clusters as in Table 1. Each cluster is a set of items as grouped by the k-medoid algorithm. Hence the items are close one another according to the similarity measure σ and intertwined by common regulatory effects enhanced by the random forest.

As we may see from the table, the sizes of the mRNA clusters are quite different. *Per se*, this is not a drawback, but simply an image of the data. We removed from our considerations 3 clusters composed of a single item, since not interesting for our purposes.

2. A Hausdorff distance matrix between the clusters of the two (mRNA and miRNA) families (see Table 2).

Looking at the histograms of these values in Fig. 5 left, we clearly identify two groups of (rescaled) distances, the ones below 0.030, let's call them small distances, and the ones up this threshold, the large distances. These groups are definitely separated, as a sharpening of the source histogram of the distances between the single miRNA-mRNA pairs represented on the right of the above graph.

Moreover, the 3D plot of the values of Table 2 highlights that the difference is made by the miRNA hand of the pairs (see Fig. 6_upper).

3. A sorted list of modules, as an immediate synthesis of the above matrix. Fig. 6_center marks with gray circles the modules with distance less than 0.25. They refer to miRNA clusters $n^\circ = 3$ and 5. Inside them, red circles mark the three modules with lowest distance, that will be invested by further considerations under a biological perspective in the next section.

Actually, the notion of distance could be further elaborated. In Fig. 6_lower we represent an analogous 3D plot where we normalize the Hausdorff distance by a factor taking into account the width of the compared clusters. Namely, we assume this factor as the inverse of the square root of the number of involved distances between miRNA-mRNA pairs. With this normalization the Hausdorff distance landscape changes, and we mark with blue square the 9 closest pairs in Fig. 6_center.

5. Discussion

To have an early evaluation of the obtained results, in Table 3 we considered the nine red-circled modules of Fig. 6_center (the closest modules), jointly with nine modules showing the highest Hausdorff distance (the furthest modules), and checked some elementary statistical properties.

As mentioned in the introduction, modules are a way of discovering joint regulatory actions of miRNAs on groups of mRNAs, like for the cliques in Fig. 1. To this aim we exploit the joint information of binding motif and expression. Namely, we use the $Y \times X$ map matrix and expression associations of miRNA pairs and mRNA pairs to *separately* group the two genomic players into clusters. Then we look for the above cliques by associating clusters from the two groups on the basis of their Hausdorff distance computed over the expression associations of mixed miRNA-mRNA pairs.

Willing to check the effectiveness of this procedure, a first indicator we adopted is the rate of targeted pairs inside a module. Looking at the more crowded modules, namely the cluster pairs ({2, 7}, {2, 1}, {2, 13}) among the closest modules and ({4, 14}, {4, 3}, {4, 0}, {1, 11}) among the furthest ones, we reckon a high rate of targeted pairs, that is generally higher in the closest module (excluding module ({2, 1}) that is over-sized). Rather, our attention is drawn by the small modules of the first group (namely the groups ({3, 3}, {3, 18}, {3, 19}) that are populated by an almost vanishing number of targeted pairs and no differentially expressed miRNAs. The first feature may pave the way to the discover of new miRNA targets, by calling either for indirect targeting instances, like those mentioned by Plotnikova et Al. [26], or for other binding mechanisms, like the combinatorial binding [24] that haven't yet been considered. The second features call for exploration of regulatory phenomena that are normally omitted. Though with a lower degree, these features characterize also the intermediate modules ({5, 3}, {5, 18}, {5, 19}) within the closest modules, but not the analogous modules of the second group.

To support these hints we mention the relations of the mRNAs GEMIN5 and EXOC8 in relation with the miRNAs hsa-miR-584d, hsa-miR-99a-5p and hsa-miR-145-5p, respectively in the closest modules {3, 19}, {5, 19} and the twelfth furthest {4, 19} module. While the sole targeting of both genes in the map matrix $Y \times X$ is uniquely declared on the part of hsa-miR-145-5p, which *per se* is not differentially expressed, biological evidence shows a regulatory interaction between those genes and the miRNA of the closest modules. In fact:

1. A study on $CD2^+$ T lymphocytes has shown that the gene GEMEN5 was significantly differentially upregulated, while hsa-miR-584d and

hsa-miR-99a-5p were differentially upregulated and downregulated respectively [13].
2. An analogous study has shown that miR-99a was strongly downregulated in breast tumor and EXOC8 was significantly upregulated [22].
3. MalaCards Disease Associated database [29] and HMDD miRNA Disease associated database [14] respectively denote, though separately, that breast cancer is a common disease for both the two genes and the three miRNAs.

In conclusion, by adopting a holistic strategy, in this paper we introduce a procedure to discover new miRNA-mRNA interactions that would be omitted in the common literature. The procedure derives from a general strategy and uses standard tools that are properly devised in order to exploit new metrics to be implemented on HPC utilities; the latter are applied to standard repositories of biological data. This work has been carried out mainly from a data analytics perspective. So, while further elaborations from a biological perspective are in order to enhance the effectiveness of the procedure, the transversality of the diseases to which the discovered interactions refer envisages its suitability in a wide spectrum of pathologies where miRNA-mRNA interactions play a relevant role.

CRediT authorship contribution statement

Ghada Shommo: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Preparation, Visualization, Investigation. **Bruno Apolloni:** Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] https://www.news-medical.net/life-sciences/microrna-profiling.aspx.
[2] M.F. Barnsley, H. Rising, Fractals Everywhere, Morgan Kaufmann, 1993.
[3] Nicolas Basalto, Roberto Bellotti, Francesco Carlo, Paolo Facchi, Ester Pantaleo, Saverio Pascazio, Hausdorff clustering, 046112, Physical review. E, Statistical, nonlinear, and soft matter physics 78 (2008) 11.
[4] Purnima Bholowalia, Arvind Kumar, Article: Ebk-means: a clustering technique based on elbow method and k-means in wsn, Int. J. Comput. Appl. 105 (9) (November 2014) 17–24.
[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth and Brooks, 1984.
[6] Kenneth Bryan, Marta Terrile, Isabella M. Bray, Raquel Domingo-FernandÃz, Karen M. Watters, Jan Koster, Rogier Versteeg, Raymond L. Stallings, Discovery and visualization of miRNA mRNA functional modules within integrated data using bicluster analysis, e17–e17, Nucleic Acids Res. 42 (3) (2013) 12.
[7] Xinqing Dai, Lizhong Ding, Hannah Liu, Zesheng Xu, Hui Jiang, Samuel K. Handelman, Yongsheng Bai, Identifying interaction clusters for mirna and mrna pairs in tcga network, Genes 10 (9) (2019).
[8] Glenn De'ath, Multivariate regression trees: a new technique for modeling species-environment relationships, Ecology 83 (4) (2002), 11051117.
[9] Reinhard Diestel, Graph Theory (Graduate Texts in Mathematics), Springer, August 2005.
[10] Jun Ding, Haiyan Hu, Xiaoman Li, SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data, e35–e35, Nucleic Acids Res. 42 (5) (2013) 12.
[11] Harsh Dweep, Carsten Sticht, Priyanka Pandey, Norbert Gretz, Mirwalk - database: prediction of possible mirna binding sites by "walking" the genes of three genomes, J. Biomed. Informatics 44 (5) (2011) 839–847.
[12] M.R. Garey, D. S. Johnson Computers, Intractability, A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences), W. H. Freeman, 1979.
[13] Yevgeniy A. Grigoryev, Sunil M. Kurian, Traver Hart, Aleksey A. Nakorchevsky, Caifu Chen, Daniel Campbell, Steven R. Head, John R. Yates, Daniel R. Salomon, Microrna regulation of molecular networks mapped by global microrna, mrna, and protein expression in activated t lymphocytes, J. Immunol. 187 (5) (2011) 2233–2243.
[14] Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, Qinghua Cui, HMDD v3.0: a database for experimentally supported human microRNAD disease associations, D1013–D1017, Nucleic Acids Res. 47 (D1) (2018) 10.
[15] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, D. Yasmin, Beazerâ Barclay, Kristen J. Antonellis, Uwe Scherf, P. Terence, Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2) (2003) 249–264.
[16] V. Jayaswal, M. Lutherborrow, D.D. Ma, and H.Y. Yee. Identification of microrna-mrna modules using microarray data. BMC Genom., 12(138), 2011.
[17] Je-Gun Joung, Kyu-Baek Hwang, Jin-Wu Nam, Soo-Jin Kim, Byoung-Tak Zhang, Discovery of microrna-mrna modules via population-based probabilistic learning, Bioinformatics (Oxford, England) 23 (9) (May 2007), 11411147.
[18] B. Liu, J. Li, A. Tsykin, Discovery of functional mirna-mrna regulatory modules with computational methods, J. Biomed. Inf. 42 (4) (2009) 685–691.
[19] B. Liu, J. Li, A. Tsykin, L. Liu, A.B. Gaur, G.J. Goodall, Exploring complex mirna-mrna interactions with bayesian networks by splitting-averaging strategy, BMC Bioinf. 10 (2009), https://doi.org/10.1186/1471-2105-10-408.
[20] Bing Liu, Lin Liu, Tsykin Anna, Gregory J. Goodall, Jeffrey E. Green, Min Zhu, Chang Hee Kim, Jiuyong Li, Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation, Bioinformatics 26 (24) (2010) 3105–3111.
[21] K.J. Livak, T.D. Schmittgen, Methods (San Diego, Calif.) 25 (4) (2001), 402408.
[22] Xinghua Long, Yu Shi, Ye Peng, Juan Guo, Qian Zhou, Yueting Tang, Microrna-99a suppresses breast cancer progression by targeting fgfr3, Front. Oncol. 9 (2020) 1473.
[23] Fionn Murtagh, Pedro Contreras, Algorithms for hierarchical clustering: an overview, WIREs Data Mining and Knowledge Discovery 2 (1) (2012) 86–97.
[24] R. Murugan, Theory on the Mechanisms of Combinatorial Binding of Transcription Factors with Dna, arXiv: Subcellular Processes, 2016.
[25] X. Pan, A. Wenzel, .L.J. Jensen, J. Gorodkin, Genome-wide identification of clusters of predicted microrna binding sites as microrna sponge candidates, PLoS One 13 (8) (2018).
[26] Olga Plotnikova, Ancha Baranova, Mikhail Skoblov, Comprehensive analysis of human microrna-mrna interactome, Front. Genet. 10 (2019) 933.
[27] Yanjun Qi, Random forest for bioinformatics, in: Ensemble Machine Learning, Springer, 2012.
[28] Frederik Questier, Raf Put, Danny Coomans, Beata Walczak, Yvan Vander Heyden, The use of cart and multivariate regression trees for supervised and unsupervised feature selection, Chemometr. Intell. Lab. Syst. 76 (2005) 45–54.
[29] Noa Rappaport, Michal Twik, Inbar Plaschkes, Nudel Ron, Tsippi ÊStein, Jacob Levitt, C. Moran Gershoni, Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search, D877–D887, Nucleic Acids Res. 45 (D1) (2016) 11.
[30] C.D.S. Kirkpatrick, Gelatt Jr., M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[31] U. Sabbagh, S. Mullegama, G.J. Wyckoff, Identification and evolutionary analysis of potential candidate genes in a human eating disorder, Biomed. Res. 2016 (2016), https://doi.org/10.1155/2016/7281732.
[32] Yuanyuan Xiao, Mark R. Segal, Identification of yeast transcriptional regulation networks using multivariate random forests, PLoS Comput. Biol. 5 (6) (2009).
[33] Xiao-Tong Yuan, Bao-Gang Hu, Ran He, Agglomerative mean-shift clustering, IEEE Trans. Knowl. Data Eng. 24 (2) (2012) 209–219.