BicNET: Efficient Biclustering of Biological Networks to Unravel Non-Trivial Modules

Rui Henriques and Sara C. Madeira

Inesc-ID and CSE Dep., Instituto Superior Técnico, Universidade de Lisboa
{rmch,sara.madeira}@tecnico.ulisboa.pt

Abstract. The discovery of dense biclusters in biological networks received an increasing attention in recent years. However, despite the importance of understanding the cell behavior, dense biclusters can only identify modules where genes, proteins or metabolites are strongly connected. These modules are thus often associated with trivial, already known interactions or background processes not necessarily related with the studied conditions. Furthermore, despite the availability of biclustering algorithms able to discover modules with more flexible coherency, their application over large-scale biological networks is hampered by efficiency bottlenecks. In this work, we propose BicNET (Biclustering NETworks), an algorithm to discover non-trivial yet coherent modules in weighted biological networks with heightened efficiency. First, we motivate the relevance of discovering network modules given by constant, symmetric and plaid biclustering models. Second, we propose a solution to discover these flexible modules without time and memory bottlenecks by seizing high efficiency gains from the inherent structural sparsity of networks. Results from the analysis of protein and gene interaction networks support the relevance and efficiency of BicNET.

1 Introduction

The increasing precision and completeness of biological networks from diverse organisms provide an unprecedented opportunity to understand the organization and dynamics of the cell [2]. In particular, the discovery of functional network modules has been largely used to characterize, discriminate and predict biological functions [29, 25, 2, 28]. The task of discovering such modules can be mapped into the discovery of coherent regions in weighted graphs, where nodes represent the molecular units (typically genes, proteins or metabolites) and the edges' weights represent the strength of the interactions between the biological molecules. In this context, a large focus has been placed on the identification of dense regions [12, 9, 11, 1], where each region is given by a statistically significant set of highly interconnected nodes. In recent years, a high number of biclustering algorithms has been proposed to discover dense regions from (bipartite) graphs by mapping them as adjacency matrices and searching for dense submatrices [1, 9, 3, 25, 22]. A bicluster is then given by two subsets of strongly connected nodes. Despite the effectiveness of biclustering to model local interactions, the focus on dense regions comes with key drawbacks. First, such regions are usually associated with either trivial or already well-known putative modules. Second, the weights of the interactions associated with less studied genes, proteins and metabolites have lower confidence (with penalizations highly dependent on the organism under study) and may not reflect the true role of these molecular interactions in certain cellular processes [31]. In particular, the presence of (well-studied) regular/background cellular processes may mask the discovery of sporadic or less-trivial processes.

Although many biclustering algorithms are able to find flexible coherencies in (adjacency) matrices [23], two major challenges have been preventing their application to biological networks. First, the generalized lack of understanding on the relevance and biological meaning of network modules with flexible coherency (given by plaid models, for example). Second, the hard combinatorial nature of biclustering regions with flexible coherency, together with the high dimensionality of matrices derived from biological networks are often associated with memory and time bottlenecks, and/or undesirable restrictions on the structure and quality of biclusters. This work aims to answer these problems by: 1) pinpointing the biological relevance of modeling non-dense regions in a network, and 2) enabling the efficient learning of flexible biclustering models from large biological networks.

To address these challenges we propose the algorithm BicNET (Biclustering NETworks). BicNET integrates contributions from pattern-based biclustering algorithms [14, 15] for the exhaustive discovery of biclusters with parameterizable coherency and quality, and adapts their data structures and searches to explore efficiency gains from the inherent sparsity of biological networks. Furthermore, we motivate the relevance of finding non-dense yet coherent modules and provide a meaningful analysis of BicNET's outputs. Results gathered from synthetic and real data show: the relevance of the proposed efficiency principles for biclustering large (possibly dense) networks, and the effectiveness of BicNET to discover a complete set of non-trivial yet coherent and biologically significant modules.

The paper is organized as follows. *Section 2* provides background on the target task of modeling functional modules given by regions with flexible coherency criteria and surveys major contributions from related work. *Section 3* proposes the BicNET algorithm. *Section 4* provides empirical evidence for the relevance of BicNET to unravel non-trivial yet relevant modules in synthetic and real networks. Finally, we draw conclusions and highlight directions for future work.

2 Background

Biclustering can be applied to different types of networks: homogeneous networks, given for instance by protein-protein interactions (PPI) and gene interactions (GI); and heteregeneous networks, capturing interactions between distinct molecular entities (proteins, protein complexes, metabolites, genes, etc.), between host and viral molecules, or between biological entities and certain terms/properties. These networks can be mapped into (bipartite) graphs for the subsequent discovery of highly interconnected regions associated with modules.

Definition 1. *Given a weighted bipartite graph with two sets of nodes* $X=\{x_1,..,x_n\}$ *and* $Y=\{y_1,..,y_m\}$*, and interactions* $a_{ij}\in\mathbb{R}$ *relating nodes* x_i *and* y_j*,* ***biclustering*** *aims to find a set of biclusters* $\mathcal{B}=\{B_1,..,B_m\}$*, where each bicluster* $B_k=(I,J)$ *is a subgraph (module) given by two subsets of nodes,* $I\subseteq X \wedge J\subseteq Y$*, satisfying specific criteria of* coherency*,* quality*, and* significance*.*

This task can be solved with traditional biclustering on real-valued matrices by mapping the bipartite graph into an adjacency matrix, where rows and columns are given by the nodes and the values by the weighted interactions. In this case, subsets of rows and columns define a bicluster associated with a network module with coherent interactions. The *structure* of a set of biclusters is defined by their number, size and positioning. Flexible structures are characterized by an arbitrary-high number of (possibly overlapping) biclusters. The *coherency* of a bicluster is defined by the observed correlation of values. Definition 2 introduces dense, constant, symmetric and plaid coherencies. The *quality* of a bicluster is defined by the type and amount of tolerated noise. The statistical *significance* of a bicluster determines the deviation of its probability of occurrence from expectations.

Definition 2. *Let the elements in a bicluster $a_{ij} \in (I,J)$ have specific coherency. A bicluster is* ***dense*** *when the average strength of its interactions, $\frac{1}{|I||J|}\Sigma_{i\in I}\Sigma_{j\in J}|a_{ij}|$, is significantly high. A* ***constant*** *coherency is observed when $a_{ij}=k_j$ where k_j is the expected strength of interactions between nodes in I and y_j node from J. In the presence of* symmetries*, $a_{ij}=k_j c_i$ where $c_i \in \{-1,1\}$. A* ***plaid*** *coherency considers cumulative contributions on the elements where biclusters/subgraphs overlap.*

Related Work on Biclustering Biological Networks. A large number of algorithms has been proposed to find modules in unweighted and/or weighted graphs mapped from homogeneous and/or heterogeneous biological networks [25, 6, 29]. In unweighted graphs, clique detection with Monte Carlo optimization [30], probabilistic motif discovery [5] and clustering on graphs [6] have been, respectively, applied to discover modules in PPIs (yeast), GIs (E. coli) and metabolic networks. In unweighted bipartite graphs, the densest regions correspond to bicliques. Bicliques can be efficiently mined using density-constrained biclustering [8], Motzkin-Straus optimization [11], formal concepts and pattern-based biclustering [3, 25, 34, 22]. In weighted graphs, the density of a module is given by the average strength of interactions. Strength is either determined by a measure of confidence (when it is predicted from literature or diverse data sources) or by the functional correlation between nodes (when it is derived from experimental data). Densely weighted modules have been discovered with betweenness-based partitioning [6], graph flow-based clustering [27] and several biclustering approaches, including SAMBA [32], multi-objective searches [24] and pattern-based biclustering [10, 9, 1]. The application of these methods over homogeneous and viral-host PPIs show that protein complexes largely match the found modules [24, 6, 27].

The discovery of dense network modules has been largely accomplished with pattern-based biclustering algorithms [9, 10, 3, 25, 34, 22, 1] due to their intrinsic ability to exhaustively discover flexible structures of biclusters. Frequent patterns in discrete networks can be mapped[1] as biclusters with specific coherency

[1] Let $\mathcal{L}$ be a finite set of items, and P an itemset $P \subseteq \mathcal{L}$. A discrete matrix D is a set of transactions in $\mathcal{L}$, $\{P_1,..,P_n\}$. Let the *coverage* Φ_P of an itemset P be the set of transactions in D in which P occurs, $\{P_i \in D \mid P \subseteq P_i\}$, and its *support* sup_P be the coverage size, $|\Phi_P|$. Given D and a minimum support θ, the *frequent itemset mining* task aims to compute: $\{P \mid P \subseteq \mathcal{L}, sup_P \geq \theta\}$.
Given D, let a matrix A be the concatenation of D elements with their column indexes. Let Ψ_P of an itemset P in A be its indexes, and Υ_P be its original items in $\mathcal{L}$. A set of *biclusters* $\cup_k(I_k,J_k)$ can be derived from frequent itemsets $\cup_k P_k$ by mapping $(I_k,J_k)=(\Phi_{P_k},\Psi_{P_k})$ to compose constant biclusters with coherency across rows ($(I_k,J_k)=(\Psi_{P_k},\Phi_{P_k})$ for column-coherency) with pattern Υ_P.

strength determined by the number of symbols (ranges of weights) assigned to the interactions. In unweighted graphs, closed frequent itemset mining and association rule mining were applied to study interactions between proteins and protein complexes in yeast proteome network [34] and between HIV-1 and human proteins [25, 22]. More recently, association rules were also used to obtain a modular decomposition of positive and negative GIs ($a_{ij}\in${-1,0,1}) [3]. In weighted graphs, Dao et. al [10] and Atluri et. al [1] relied on the loose antimonotone property of density to propose weight-sensitive pattern mining searches. DECOB [9], originally applied to PPIs and GIs from human and yeast, uses an additional filtering step to output of non-similar modules only.

Some of these works have been extended to discover discriminative modules, often referred as multigenic markers, for classification tasks such as function prediction [10, 29, 22]. Network-based (bi)clustering methods for function prediction have been comprehensively reviewed by Sharan et. al [29].

Related Work on Biclustering Modules with Flexible Coherency. Although the state-of-the-art is focused on the discovery of dense network modules, slight variants of this coherency have been proposed [32, 19, 1]. Despite the large availability of biclustering algorithms able to find biclusters with flexible coherency [23], empirical evidence shows that they are not prepared to deal with the sparsity and/or high-dimensionality of adjacency matrices mapped from networks. A first attempt towards this end was presented by Tomaino et al. [33] for small networks.

3 Solution

In what follows, we first show how biclustering can be applied to discover coherent modules following constant, symmetric and plaid models, possibly containing noisy and missing interactions. Second, we extend pattern-based searches to optimally handle the inherent structural sparsity of biological networks.

3.1 Network Modules with Flexible Coherency

Biclustering Weighted Graphs. For an effective application of state-of-the-art biclustering algorithms to (weighted) graphs derived from biological networks, two principles should be satisfied. First, the weighted graph should be mapped into a minimal bipartite graph. In heterogeneous networks, multiple bipartite graphs are created (each with two disjoint sets of nodes with heterogeneous interactions). The minimality requirement can be satisfied by identifying subsets of nodes with cross-set interactions but without intra-set interactions to avoid unnecessary duplicated nodes in the disjoint sets of nodes (see Fig.1). This is essential to avoid the generation of large bipartite graphs and subsequent very large matrices.

Second, when targeting non-dense coherencies, two real-valued adjacency matrices need to be derived from the bipartite graph (a matrix with rows and columns mapped from the disjoint sets of nodes and its transpose) for an exhaustive space exploration. This is different from using all nodes as rows and columns in a single matrix and then filling the upper and lower triangular matrices, which can can lead to inconsistencies when a bicluster has elements from both the upper and lower triangular matrices. Also, the larger size and density of such matrix can significantly hamper the efficiency of the biclustering task. The few attempts to find non-dense biclusters in biological networks fail to satisfy this principle [33], thus delivering incomplete and often inconsistent solutions.

Pattern-based Biclustering. Under the satisfaction of the previous principles, a wide-range of biclustering algorithms can be applied to discover modules with flexible coherencies [23]. Yet, to our knowledge, only pattern-based biclustering [14–16] is able to guarantee an exhaustive yet efficient discovery of flexible structures of biclusters with parameterizable coherency and quality criteria. This provides the necessary context to measure the relevance and impact of discovering modules with non-dense coherency and noise-tolerance. In particular, we rely on BicPAM and BiP algorithms [15, 13]. These algorithms, respectively, use frequent itemset mining and association rule mining to find biclusters with constant/symmetric and plaid coherencies. Furthermore, they integrate the dispersed contributions from previous pattern-based algorithms and address some of their limitations, providing key principles to surpass discretization problems (by introducing the possibility to assign multiple symbols to a single element) and robustly handle noise and missing values. Fig.1 provides a view on how transactions can be derived from (heterogeneous) biological networks for the discovery of constant modules (see [15] for details on the itemization, mining and postprocessing steps).

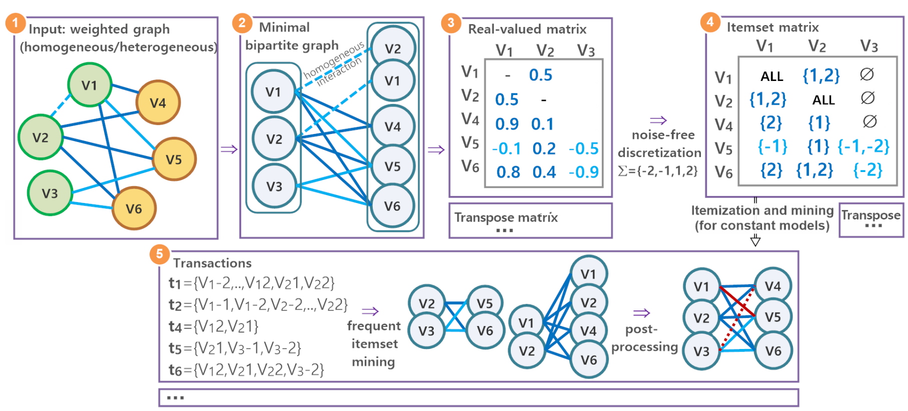

Fig.1: Pattern-based biclustering of (heterogeneous) biological networks.

Constant Model. Given a bicluster defining a module with coherent interactions between two sets of nodes, the constant coherency (Definition 2) implies that the nodes in one set show a single type of interaction with the nodes in the remaining set. Illustrating, consider a set of interactions between genes and proteins, where their absolute weight defines the strength of the association and their sign determines whether the association corresponds to activation or repression mechanisms. The constant model guarantees that when a gene is associated with a group of proteins, it establishes the same type of interaction with all these proteins (such as heightened activation of the transcription of a complex of proteins). When analyzing the transposed matrix (by switching the disjoint sets of the bipartite graph), similar relations can be observed: a protein coherently affects a set of genes (softly repressing their expression, for example). The constant model can also disclose relevant interactions between homogeneous groups of genes, proteins and metabolites. Fig.2 provides an illustrative constant module.

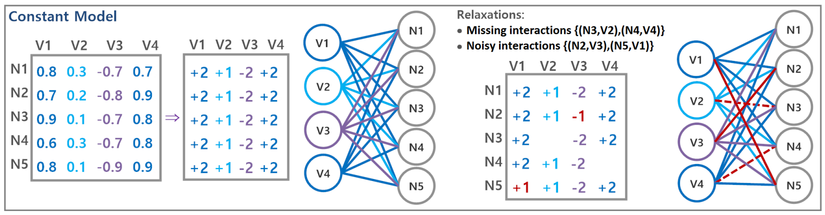

Fig.2: Biclustering (noise-tolerant) modules with the constant model.

The constant model can be further applied to networks with qualitative interactions capturing distinct types of regulatory relations, such as *binds*, *activates* or *enhances* associations, common in a wide-variety of PPIs [22, 25].

The constant model is essential to guarantee that molecular units with non-necessarily high (yet coherent) influence on another set of molecular units are not excluded. The constant coherency is in general more flexible than the dense coherency, leading to the discovery of larger modules. The exception is when the dense coherency is not given by highly weighted interactions, but instead by all interactions independently of their weight (extent of interconnected nodes).

Symmetric Model. The presence of symmetries is key to simultaneously capture activation and repression mechanisms associated with the interactions of a single node [15]. The symmetric model introduces a new degree of flexibility by enabling the discovery of more complex regulatory modules, where a specific gene/protein may show symmetric regulatory behavior according to the expected pattern, yet still respect the observed coherency. Fig.3 illustrates the symmetric model, where rows with symmetries are identified with dashed lines.

Plaid Model. The plaid assumption [13] is essential to describe overlapping regulatory behavior associated with cumulative effects in the strength of interactions between nodes that appear in multiple functional modules. Illustrating, consider that two genes interact in the context of multiple biological processes, a plaid model can consider their cumulative effect on their interaction's weights (based on the expected weight associated with each active process). This is also valid for the regulatory influence between proteins and for heterogeneous networks. The plaid assumption of GIs and PPIs also provides insights on the network topology and molecular functions, revealing hubs and core interactions (based on the amount of overlapping interactions), and between- and within-pathway interactions (based on the interactions inside and outside of the overlapping areas). Fig.3 illustrates a plaid model associated with two overlapping modules. These modules could not be discovered without a plaid assumption.

Handling Noisy and Missing Interactions. An undesirable restriction of exhaustive searches for dense modules is that they may exclude relevant nodes associated with a bicluster if those nodes do not interact with all of the nodes in one subset of nodes from the bicluster. Understandably, meaningful modules with missing interactions are common since the majority of existing biological networks are still largely incomplete. Pattern-based biclustering is able to recover missing

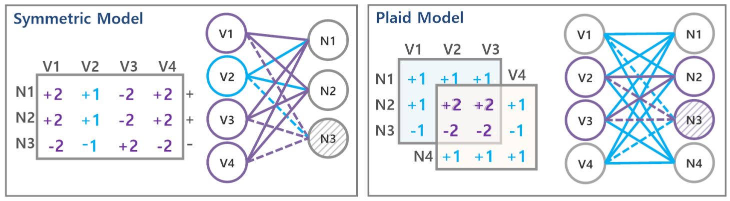

Fig.3: Biclustering modules with the symmetric and plaid models.

interactions recurring to well-established and efficient postprocessing procedures (based on the merging and extension of the discovered modules) [15].

Furthermore, the scoring scheme of interactions might be prone to experimental noise, preprocessing biases and structural noise (particularly common for less studied and stable genes or proteins), not always reflecting the true interactions. Pattern-based biclustering also allows the assignment of multiple symbols to specific interactions [15], thus avoiding the exclusion of noisy interactions (see Fig.1). Although default parameterizations are provided to guarantee an adequate tolerance to noise, the level of sparsity and noise of the discovered modules can be parametrically controlled using thresholds based on quality expectations. Fig.2 shows an illustrative coherent module with corrections associated with missing interactions (red dashed lines) and noisy interactions (red continuous lines).

3.2 BicNET: Efficient Biclustering of Biological Networks

Understandably, the task of discovering modules with the introduced coherencies is more complex than finding dense modules (complexity discussed in [15]). Empirical evidence shows that state-of-the-art biclustering algorithms are only scalable for biological networks up to a few hundreds of nodes (see Results). Nevertheless, a key property distinguishing biological networks from gene expression or clinical data is their underlying sparsity. Illustrating, some of the densest PPI and GI networks from well-studied organisms still have a density below 5% (ratio of interconnected nodes after excluding nodes without interactions). While traditional biclustering depends on operations over matrices, pattern-based biclustering algorithms are prepared to mine transactions of varying length. This property makes pattern-based biclustering able to exclude missing interactions from searches and thus surpass memory and efficiency bottlenecks. Based on this observation, we propose BicNET (**BiC**lustering Biological **NET**works), a pattern-based biclustering algorithm for the discovery of network modules with non-trivial coherencies and robustness to noise. Additionally, BicNET relies on the following principles to explore further efficiency gains.

We propose a new data structure to efficiently preprocess data: an array, where each position (node from a disjoint set in the bipartite graph) has a list of pairs, each pair representing an interaction (corresponding node and the interaction weight). Discretization and itemization procedures are performed by linearly scanning this structure three times. Thus, their time and memory complexity is linear on the number of interactions.

Pattern-based searches commonly rely on bitset vectors due to the need to retrieve not only the frequent patterns but also their supporting transactions in order to compose biclusters. However, bitset vectors are costly in terms of memory, and the associated intersection operations are computationally expensive for large-scale networks. For this reason, we rely on the recently proposed F2G miner [17] and on revised implementations of Eclat and Charm miners where diff-sets are used to address the bottlenecks of bitsets. These pattern-based searches guarantee an efficient discovery of constant, symmetric and plaid models.

Furthermore, the underlying pattern mining searches of BicNET are dynamically selected based on the properties of the network to optimize their efficiency. Horizontal versus vertical data formats [15] are selected based on the ratio of rows and columns from the mapped matrix. Apriori (candidate generation) versus pattern-growth (tree projection) searches [15] are selected based on network density (pattern-growth searches are preferable for dense networks). We also push the computation of similarities between all pairs of biclusters (the most expensive postprocessing procedure) into the mining step by checking similarities with distance operators on a compact data structure to store the frequent patterns.

4 Results and Discussion

Results are organized as follows. First, we compare the performance of BicNET against state-of-the-art biclustering algorithms using synthetic networks. Second, we use BicNET for the analysis of large-scale PPI and GI networks to show the relevance of discovering modules with flexible coherencies and parameterizable levels of noise and sparsity. BicNET is implemented in Java (JVM v1.6.0-24). Experiments were computed using an Intel Core i5 2.30GHz with 6GB of RAM.

Synthetic Data. Networks with planted biclusters were generated respecting the commonly observed topological statistics of biological networks. Variables:

- number of nodes, density and distributions of the weight (positive and negative ranges revealing the interaction strength);
- degree of noisy and missing interactions (from 0% to 20%).
- number, size (Uniform distribution on the number of nodes), shape (imbalance on the size of the disjoint sets of each subgraph), overlapping, and coherency (dense, constant, symmetric and plaid) of the planted biclusters:

	Network nodes (10% density)					Network density (2000 nodes)			
	200	500	1000	2000	10000	1%	5%	10%	25%
♯ Hidden modules	5	10	15	20	30	3	5	10	20
♯ Nodes per module	[20,30]	[30,40]	[40,50]	[50,70]	[100,140]	[50,70]	[50,70]	[50,70]	[50,70]
% Interactions in modules	19,5%	12,2%	7,6%	4,5%	1,1%	22,5%	9,0%	4,5%	2,3%

Real Data. We used four biological networks: GIs in yeast from DryGIN [21] and STRING v10 [31] databases, and two licensed PPIs in human and E. coli from STRING v10 [31] database. The scores in these networks show the expected strength of influence/physical interaction between genes/proteins (see Table 1 for statistics).

Performance Metrics. Given the set of planted modules $\mathcal{H}$ in a synthetic network, the accuracy of the retrieved modules $\mathcal{B}$ is given by two match scores (1): $MS(\mathcal{B},\mathcal{H})$ defining the extent to what the found biclusters cover the hidden

Type	Organism	♯Nodes	♯Interactions	Density	Notes on the weight of interactions
GI	Yeast	4455	191309	1.0%	Weights (65% negative) from double-mutant arrays [21].
GI	Yeast	6314	423335	1.1%	Known and predicted associations benchmarked
PPI	E. Coli	8428	3293416	4.6%	from multiple data sources and text mining,
PPI	Human	19247	8548002	2.3%	and combined through an integrative score [31].

Table 1: Biological networks used to assess the relevance and efficiency of BicNET.

biclusters (*completeness*), and $MS(\mathcal{H},\mathcal{B})$ reflecting how well the hidden biclusters are recovered (*precision*). We present the average of matches collected from 10 instantiations of synthetic networks. These accuracy criteria surpass the problems of Jaccard matches (only focused on one of the two subsets of nodes at a time [15]) and RNIA (loose matching criteria [15]). Efficiency and significance are used to complement this analysis.

$$\mathbf{MS}(\mathcal{B},\mathcal{H}) = \frac{1}{|\mathcal{B}|}\Sigma_{(I_1,J_1)\in\mathcal{B}} max_{(I_2,J_2)\in\mathcal{H}} \sqrt{\frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}\frac{|J_1 \cap J_2|}{|J_1 \cup J_2|}} \quad (1)$$

4.1 Results on Synthetic Data

Fig.4 compares the efficiency of BicNET with state-of-the-art biclustering algorithms with flexible coherence criteria using networks with varying size and density and planted modules with constant coherency. We selected FABIA[2] [18], ISA [20], xMotifs [26], CC [7] and OPSM [4] to discover modules with flexible coherency. BicNET shows heightened efficiency levels. Understandably, as most of the remaining algorithms are only prepared to analyze (non-sparse) matrices, they show efficiency bottlenecks for even small networks. Furthermore, the majority is not able to accurately recover the planted modules as they cannot interpret missing interactions. Although SAMBA [32] and some pattern-based biclustering algorithms, such as BiMax and DECOB [25, 9], are able to discover dense models efficiently, they are not prepared to discover modules with alternative coherence criteria.

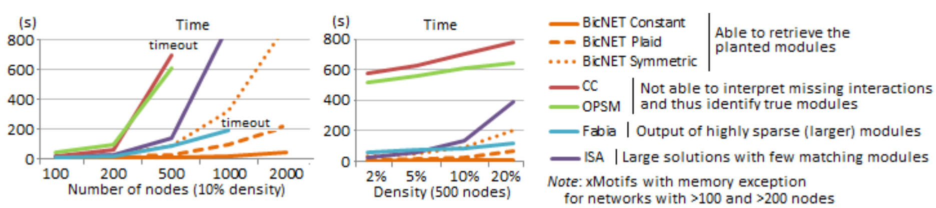

Fig.4: Efficiency of flexible biclustering algorithms to discover constant modules in synthetic networks with varying size and density.

Fig.5 zooms-in on the performance of BicNET by quantifying the efficiency gains in memory and time from using adequate data structures (replacing the need to use matrices) and searches (replacing the need to rely on bitset vectors). It also shows that the cost of assigning multiple symbols per interaction are moderate, despite resulting in an increased network density.

[2] Sparse prior equation with decreasing sparsity until able to retrieve a non-empty set of biclusters

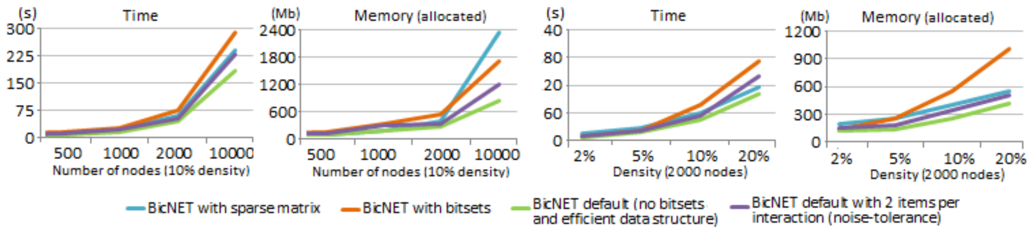

Fig.5: Efficiency gains of BicNET when using sparse data structures, pattern mining searches providing robust alternatives to bitset vectors, and noise handlers.

Fig.6 compares the performance of BicNET with peer algorithms for discovering dense network modules (hypercliques) in the presence of noisy and missing interactions. This analysis clearly shows that existing pattern-based searches for hypercliques have no tolerance to errors since their accuracy rapidly degrades for an increased number of planted noisy/missing interactions. Thus, they are not able to deal with the natural incompleteness and scoring uncertainty associated with biological networks. On the other hand, the observed accuracy levels of BicNET demonstrate its robustness to noise (validating the importance of assigning multiple ranges of weights for some interactions) and to missing interactions (showing the effectiveness of BicNET's postprocessing procedures).

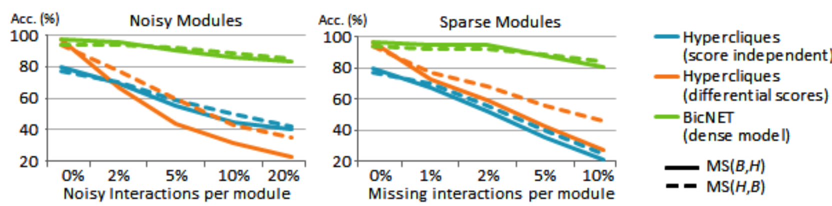

Fig.6: Accuracy of BicNET against peer pattern-based searches to discover dense modules on networks (2000 nodes, 10% density) with varying degree of noise and missings.

4.2 Results on Real Data

The biological significance of the modules discovered in real data was computed by assessing the over-representation of Gene Ontology (GO) terms with an hypergeometric test. A module is significant when its genes show enrichment for one or more terms by having a (Bonferroni corrected) p-value below 0.01. Fig.7 shows the properties of BicNET's solutions for the four biological networks in Table 1. 94% of the modules discovered in DRYGIN's yeast GIs were significantly enriched. All the modules discovered in STRING's yeast GIs were significantly enriched. BicNET was able to discover the largest number of (non-similar and statistically significant) biclusters. The analysis of the enriched terms for these modules against the enriched terms found in other biclustering solutions supports the completeness, exclusivity and relevance of BicNET's solutions (Table 2). The significance of peer solutions from unweighted graphs is penalized by the inability to remove nodes with either low or non-coherent weights, while the significance of peer solutions focused on dense regions is additionally hampered by noise and discretization errors (Fig.7).

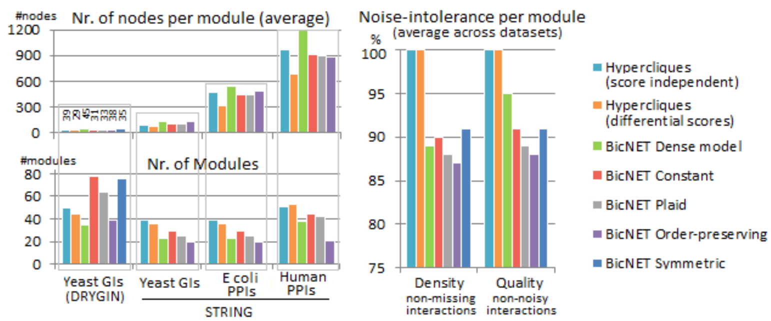

Fig.7: Properties of BicNET's solutions with varying coherency against peer pattern-based searches for dense modules (hypercliques) in networks from DRYGIN and STRING.

Table 2 shows the properties of an illustrative set of significantly enriched modules. We can observe that such biclusters could hardly be discovered by peer methods due to their non-dense coherency. All of the illustrated modules show coherent patterns of interaction between nodes combining both differential and non-differential weights. The provided modules have an average of 5 to 10% of missing interactions. BicNET is well positioned to find modules with varying size, coherency and quality. Illustrating, the constant modules D_6 and D_7 have, respectively, 23 and 47 nodes and distinct quality, being D_7 more tolerant to noisy interactions. Understandably, the number of nodes per module is naturally affected by the size and sparsity of the target network. Most of the discovered modules clearly show non-trivial yet meaningful correlations, whose relevance is pinpointed by the number of highly enriched terms after correction.

	ID	Type	♯Nodes $\lvert I\rvert\times\lvert J\rvert$	Items	♯Terms p<1E-15	Notes
DryGIN	D1	constant	18×9	{-4,..,4}	27	Module with coherent strong (-4) and soft (-1) negative interactions.
	D2	symmetric	4×9	{-3,..,3}	13	Varying levels of strong (mainly positive) interactions ({±3,±2}).
	D3	symmetric	5×6	{-2,-1,1,2}	12	Module with either all positive or negative interactions per "row"-node ({±1,±2}).
	D4	constant	7×5	{1,2}	12	Module with coherent strong (2) and soft (1) positive interactions.
	D5	symmetric	7×5	{-2,-1,1,2}	11	Module with either all positive or negative interactions per "row"-node ({±1,±2}).
	D6	constant	13×10	{-2,-1,1,2}	24	Module with mostly strong negative interactions per "row"-node.
	D7	constant	39×8	{-2,-1,1,2}	47	Noise-tolerant module with positive and negative interactions.
STRING	S1	constant	148×13	{1,2}	169	Noise-tolerant module with positive interactions of varying strength ({1,2}).
	S2	constant	80×18	{1,2,3}	98	Module with mostly of non-dense interactions ({1,2}).
	S3	constant	83×10	{1,2}	93	Module with non-dense positive interactions before postprocessing ({1}).
	S4	constant	50×20	{1,2,3}	70	Module with non-dense positive interactions ({1,2}) before postprocessing.
	S5	constant	45×31	{1,2,3}	76	Module with mostly dense interactions (weights in {2,3}).
	S6	constant	55×85	{1,2}	143	Module with mostly dense interactions ({2}).

Table 2: Exclusivity and relevance of BicNET solutions: properties of found modules.

Table 3 lists some of the enriched terms for the modules in Table 2, showing their functional coherence and role to unravel putative biological processes. Interesting, some of the identified modules are part of an additive plaid model (with in-between condition [13]). Illustrating, modules D6 and S4 share, respectively, 21% and 36% of their interactions with modules D7 and S4 under a plaid assumption. Without this assumption, only smaller modules (excluding key nodes) could be obtained, resulting in a lower enrichment of their terms.

In a concluding note, when analyzing networks derived from knowledge-based repositories and literature (such as networks from STRING [31]), the flexibility of coherence and noise-robustness is critical to deal with uncertainty and re-

	ID	Terms description (♯)	♯Terms p<1E-15	♯Nodes
DryGIN	D1	Histone modification; regulation of histones: H3-K79/H3-K4 methylation, H2B ubiquitination, etc. (5);	6	27
	D2	Gluconeogenesis; glutamate metabolic/catabolic processes (2); nicotinamide metabolism/biosynthesis (2);	6	13
	D3	Positive and negative regulation of transcription from RNA polymerase II; Invasive growth response to glucose limitation and hyperosmotic salinity response by regulating RNA polymerase II (5);	5	12
	D4	Meiotic anaphase I; activation of anaphase-promoting complex activity involved in meiotic cell cycle;	4	12
	D5	Negative reg. of phospholipid biosynthesis; lipid homeostasis; isopropylmalate and oxaloacetate transport;	4	11
	D6	Cotranslational protein targeting to membrane; protein insertion into mitochondrial membrane; protein import into peroxisome membrane; reg. sporulation; actin filament bundle assembly involved in cytokinesis;	5	25
	D7	Acetate fermentation, acetyl-CoA biosynthesis (from acetate), reg. transcription on exit from mitosis;	7	50
STRING	S1	Response to hypoxia; oxidation-dependent protein catabolic process; anaerobic respiration; age-dependent response to reactive oxygen species; cellular response to oxidative stress;	36	169
	S2	Positive & negative reg. of mitotic and nuclear cell cycle, DNA replication, budding cell apical bud growth;	16	98
	S3	Transport of aerobic e-, acetyl-CoA, vacuolar transm., amine (5); ribose phosphate & D-ribose processes (2);	22	93
	S4	Heterochromatin maintenance involved in chromatin silencing; sister chromatid segregation;	6	70
	S5	Cytoplasmic and mitochondrial translation (4); regulation of translational fidelity; ADP biosynthesis;	6	76
	S6	rRNA processing; separation, cleavage & maturation of SSU-rRNA (5); ribosomal (large subunit) biogenesis;	14	143

Table 3: Illustrative set of biologically significant BicNET's modules: description of the highly enriched terms in the modules presented in Table 2.

gions where weights may be affected due to the unbalanced focus of research studies. When analyzing networks derived from data experiments (such as GIs from DRYGIN [21]), the discovery of modules with non-necessarily strong interactions (given by the constant model, for example) can be critical to identify less-predominant (yet key) biological processes, such as the ones associated with early stages of stimulation or disease.

5 Conclusions and Future Work

This work motivates and answers the task of biclustering large-scale biological networks to discover modules with flexible yet meaningful coherency and robustness to noise. In particular, we explored the relevance of mining non-trivial modules in both homogeneous and heterogeneous networks, and assessed the limits in efficiency of existing biclustering algorithms targeting non-dense models. Combining state-of-the-art contributions on pattern-based biclustering with efficient searches on networks, we propose BicNET algorithm for the exhaustive discovery of constant, symmetric and plaid models in biological networks. Additional strategies are further incorporated to retrieve modules with noisy and missing interactions, thus addressing the limitations of the existing exhaustive searches on networks. BicNET enables the analysis of dense networks with up to 50000 nodes. Results on synthetic and real networks confirm its efficiency and relevance to discover non-trivial (yet coherent and significant) modules.

Six possible directions are identified for future work: to consider further coherencies such as order-preserving and scale factors; enhance searches with scalability principles from pattern mining (data partitioning strategies and search for approximate patterns [14]); extend the proposed contributions for the integrative mining of network and expression data; explore the relevance of the plaid model to identify and characterize hubs; enlarge the experimental analyzes towards biological molecules with yet unclear roles; and embrace predictive tasks.

Acknowledgments. This work was supported by *FCT* under the project UID/CEC/50021/2013 and the PhD grant SFRH/BD/75924/2011 to RH.

References

1. Atluri, G., Bellay, J., Pandey, G., Myers, C., Kumar, V.: Discovering coherent value bicliques in genetic interaction data. In: IW on Data Mining in Bioinformatics (2010)
2. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
3. Bellay, J., Atluri, G., Sing, T.L., Toufighi, K., Costanzo, M., et al.: Putting genetic interactions in context through a global modular decomposition. Genome Research 21(8), 1375–1387 (2011)
4. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering local structure in gene expression data: the order-preserving submatrix problem. In: RECOMB. pp. 49–57. ACM (2002)
5. Berg, J., Lässig, M.: Local graph alignment and motif search in biological networks. National Academy of Sciences 101(41), 14689–14694 (2004)
6. Chen, J., Yuan, B.: Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics 22(18), 2283–2290 (2006)
7. Cheng, Y., Church, G.: Biclustering of expression data. In: ISMB. pp. 93–103. AAAI (2000)
8. Colak, R.: Towards finding the complete modulome: density constrained biclustering. Ph.D. thesis, Simon Fraser University (2008)
9. Colak, R., Moser, F., Chu, J.S.C., Schönhuth, A., Chen, N., Ester, M.: Module discovery by exhaustive search for densely connected, co-expressed regions in biomolecular interaction networks. PLoS One 5(10), e13348 (2010)
10. Dao, P., Colak, R., Salari, Moser, Davicioni, Schönhuth, Ester, M.: Inferring cancer subnetwork markers using density-constrained biclustering. Bioinformatics 26(18), i625–i631 (2010)
11. Ding, C., Zhang, Y., Li, T., Holbrook, S.: Biclustering protein complex interactions with a biclique finding algorithm. In: ICDM. pp. 178–187 (2006)
12. Georgii, E., Dietmann, S., Uno, T., Pagel, P., Tsuda, K.: Enumeration of condition-dependent dense modules in protein interaction networks. Bioinformatics 25(7), 933–940 (2009)
13. Henriques, R., Madeira, S.: Biclustering with flexible plaid models to unravel interactions between biological processes. IEEE/ACM TCBB (2015)
14. Henriques, R., Antunes, C., Madeira, S.C.: A structured view on pattern mining-based biclustering. Pattern Recognition (2015)
15. Henriques, R., Madeira, S.: Bicpam: Pattern-based biclustering for biomedical data analysis. Algorithms for Molecular Biology 9(1), 27 (2014)
16. Henriques, R., Madeira, S.C.: Pattern-based biclustering with constraints for gene expression data analysis. In: Computational Methods in Bioinformatics and Systems Biology (EPIA-CMBSB). LNAI, Springer (2015)
17. Henriques, R., Madeira, S.C., Antunes, C.: F2g: Efficient discovery of full-patterns. In: ECML/PKDD IW on New Frontiers to Mine Complex Patterns. Springer-Verlag (2013)
18. Hochreiter, S., et al.: FABIA: factor analysis for bicluster acquisition. Bioinf. 26(12), 1520–1527 (2010)
19. Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(suppl 1), S233–S240 (2002)
20. Ihmels, J., Bergmann, S., Barkai, N.: Defining transcription modules using large-scale gene expression data. Bioinformatics 20(13), 1993–2003 (Sep 2004)
21. Koh, J.L.Y., Ding, H., Costanzo, M., Baryshnikova, A., Toufighi, K., Bader, G.D., Myers, C.L., Andrews, B.J., Boone, C.: Drygin: a database of quantitative genetic interaction networks in yeast. Nucleic Acids Research 38(suppl 1), D502–D507 (2010)
22. MacPherson, J.I., Dickerson, J., Pinney, J., Robertson, D.: Patterns of hiv-1 protein interaction identify perturbed host-cellular subsystems. PLoS Comput Biol 6(7), e1000863 (2010)
23. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM TCBB 1(1), 24–45 (2004)
24. Maulik, U., Mukhopadhyay, A., Bhattacharyya, M., Kaderali, L., Brors, B., Bandyopadhyay, S., Eils, R.: Mining quasi-bicliques from hiv-1-human protein interaction network: A multiobjective biclustering approach. IEEE/ACM TCBB 10(2), 423–435 (2013)
25. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: A novel biclustering approach to association rule mining for predicting hiv-1–human protein interactions. PLoS ONE 7(4), e32289 (04 2012)
26. Murali, T.M., Kasif, S.: Extracting conserved gene expression motifs from gene expression data. In: Pacific Symposium on Biocomputing. pp. 77–88 (2003)
27. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. PROTEINS: Structure, Function, and Bioinformatics 54(1), 49–57 (2004)
28. Segal, E., Wang, H., Koller, D.: Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19(suppl 1), i264–i272 (2003)
29. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular systems biology 3(1) (2007)
30. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. National Academy of Sciences 100(21), 12123–12128 (2003)
31. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al.: String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research p. gku1003 (2014)
32. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics 18, 136–144 (2002)
33. Tomaino, V., Guzzi, P.H., Cannataro, M., Veltri, P.: Experimental comparison of biclustering algorithms for ppi networks. In: BCB. pp. 671–676. ACM (2010)
34. Xiong, H., Heb, X.F., Ding, C., Zhang, Y., Kumar, V., Holbrook, S.R.: Identification of functional modules in protein complexes via hyperclique pattern discovery. Pac. Symp. Biocomput. (2005)