Supplementary material

CALI: A novel visual model for frequent pattern mining in protein-ligand graphs

Susana Medina G.[1]* , Alexandre V. Fassio[1], Sabrina A. Silveira [2], Carlos H. da Silveira[3], and Raquel C. de Melo-Minardi[1]

[1]Department of Computer Science, UFMG, Belo Horizonte, 31270-901, Brazil,
[2]Department of Computer Science, UFV, Viçosa, 36570-000, Brazil and
[3]Advanced Campus at Itabira, UNIFEI, Itabira, 35903-087, Brazil.

*To whom correspondence should be addressed.

1 Methods

1.1 Network metrics

Here we explain, according to (Newman, 2010), the complex network metrics that support the empirical analysis of graphs obtained with CALI model.

- Degree: the simplest centrality metric is the degree of a node, which is the number of edges connected to it. (Newman, 2010). Is also known as the number of neighbours of a given node in a graph. For example, in Figure 1 there are five hubs representing hydrogen bond interactions with high degree: the highest (78) is linked with three nodes to another one with 42, another big one with 61 and two smaller with 16 and 14.
- Node betweenness: measures the extent to which a node lies on paths between other nodes. This metric allows to identify important nodes with low degree, which also works as a bridges, joining two or more groups. The betweenness centrality of a node i is the number of all shortest paths in the graph that pass through i:

$$x_i = \sum_{st} n_{st}^i \tag{1}$$

  s and t are nodes of the graph and if there is no path between them n_{st}^i is zero.
- Edge betweenness: similarly, measures the extent to which an edge lies on paths between nodes.
- Closeness: measures the means distance from a node to other nodes. Suppose $d_{i,j}$ is the length of a shortest path between nodes i and j, then the mean shortest path distance from i to j, averaged all vertices j in the network. (Newman, 2010) uses the following formula:

$$l_i = \frac{1}{n} \sum_j d_{ij} \tag{2}$$

- Eccentricity: is the maximum distance between a node s and any reachable node t of the graph (Junker *et al.*, 2008). Distance between two nodes that are not connected is defined as infinity. In this case, eccentricity can be calculated for the graph connected components. The eccentricity centrality for s is defined by the formula:

$$ecc(s) = \frac{1}{max\{dist(s,t) : t \in V\}} \tag{3}$$

- Communicability [1]: known as subgraph centrality of a node, it is the sum of closed walks of all lengths starting and ending at the same node. Using the graph spectrum notation in the generalization proposed by (Estrada and Hatano, 2008), where $\phi_j(s)$ is the s th element of the j th orthogonal eigenvector of the adjacency matrix associated with the eigenvalue λ_j , the communicability between the nodes s and t is given by:

$$c(s,t) = \sum_{j=1}^{n} \phi_j(s)\phi_j(t)e^{\lambda_j} \tag{4}$$

[1] https://networkx.github.io/documentation

1.2 The FSM strategy

A well stablished strategy to find patterns in a dataset of graphs is the use of *frequent subgraph mining* (FSM) algorithms as, for instance, SUBDUE Cook *et al.* (1994), AGM Inokuchi *et al.* (2000), FSG Kuramochi *et al.* (2004), MoFa Borgelt *et al.* (2002), gSpan Yan *et al.* (2002), FFSM Huan *et al.* (2003) or GASTON Nijssen *et al.* (2005).

The work Silveira *et al.* (2015) instantiated the FSM paradigm and focused on discovering PLI patterns through a *transaction based* and *exact algorithm*, gSpan. *Transaction based* FSM algorithms aim to find frequent subgraphs in a collection of input graphs, called transactions, and *exact* means that the mining algorithm guarantees to find all frequent subgraphs in the input data Jiang *et al.* (2013). This strategy allows to reveal some residues and even atoms that are essential for PLI according to experimental biological data. Nonetheless, this strategy have some drawbacks that we discuss bellow.

Exact frequent subgraph mining algorithms are computationally expensive as they undertake extensive subgraph isomorphism comparisons Jiang *et al.* (2013). In addition, the number of frequent subgraphs increases exponentially with the size of the graph (for a frequent k-graph, the number of frequent subgraphs can be as large as 2^k) and, among the patterns, many can be structurally repetitive, as a frequent subgraph can have other frequent subgraphs within it Yan *et al.* (2003). Generating and analyzing the results of this data mining process is certainly a challenge.

Finding a large number of patterns, as many of them are structurally repetitive, is a problem that can be mitigated by using FSM algorithms that mine closed (CLOSECUT and SPLAT, both proposed by Yan *et al.* (2003)) or maximal (SPIN Huan *et al.* (2004) and MARGIN Thomas *et al.* (2006)) subgraphs. According to Koyutürk *et al.* (2004), in the case of biological networks, maximal frequent subgraphs are deemed to be the most interesting ones, while according to Fischer *et al.* (2004), closed patterns have some biological meaning.

The main disadvantage of FSM algorithms is that they do not provide a direct mapping from output patterns to the graphs in the input dataset. When studying PLIs, it is crucial to map computed patterns to protein and ligand atoms in the input data, which gives the information of which atoms (and its respective residues) are involved in PLI throughout a dataset of interest. To the best of our knowledge, the only way to circumvent this issue is to map the frequent subgraphs from FSM output to input data by performing subgraph isomorphism, which is, according to Bonnici *et al.* (2015), a computationally expensive (NP-complete) problem. The computational cost of frequent subgraph mining algorithms, combined with such mapping process, prevents FSM strategy from being general and scalable.

1.3 Datasets

In this section we detail how CDK2 and Ricin datasets were obtained (from Protein Data Bank (PDB) (Rose *et al.*, 2015) in August 2014). The list of PDB identifiers comprising each dataset is provided in Table 1 for CDK2 and Table 2 for Ricin.

1.3.1 CDK2

Schonbrunn and colleagues in their work (Schonbrunn *et al.*, 2013) depict how they discovered by high-throughput screening the compound 2-(allylamino)-4-aminothiazol-5-yl-(phenyl) methanone as a potent inhibitor of the human CDK2. Through the cocrystal structure (PDB id 3QQK), they could show the importance of hydrogen bonds in the binding of this compound with the ATP site. The hydrogen bonds occur between the thiazolamine moiety and the hinge region (GLU81-LEU83).

Departing from previously cited compound, the authors developed other 95 analogues by replacing systematically the flanking allyl and the phenyl moieties whereas the aminothiazole core was maintained unchanged to preserve its functionality. Thenceforth, they evaluate analogues as their inhibitory potential. Nonetheless, only 35 from these analogues had their crystal structure determined.

Besides these 35 structures, we have found another 38 related structures (totaling 73 PDB files) by searching on the PDB[2] website. These files are supposed to be discussed in another work from the same authors which is not published yet to the best of our knowledge.

1.3.2 Ricin

We searched Protein Data Bank for key words *ricin*, *ricin-like* and *ribosome inactivating protein* and obtained 136, 126 and 163 results respectively. As there was overlap among results, the total number of different PDB entries was 266.

Sequences from all 266 PDB entries were split by chain using PDBest tool (Gonçalves *et al.*, 2015) and were aligned against PDB id 2AAI (Rutenber *et al.*, 1991) chain A, which we call 2AAI.A, using an in-house implementation of Needleman-Wunsch algorithm (Needleman *et al.*, 1970). PDB entry 2AAI.A is the catalytic subunit of ricin toxin without any ligands. Those 47 structures which have 50% or more identity were taken as our initial ricin dataset.

The final step to obtain ricin dataset was to select entries which have at least one ligand, as we were interested in patterns of interactions between a protein and its ligands. So we computed probable protein-ligand interactions at atomic level using a geometric approach (which is detailed in Section *Modelling of protein-ligand interactions as graphs*) to determine ligands that were interacting with protein residues. Only ligands with seven or more atoms were considered, in a similar manner to (Pires *et al.*, 2013). This process resulted in 29 PDB chains.

2 Results and discussion

2.1 Global CALI model analysis

Global network descriptors are provided in Table 3. *Global CALI model analysis* from main paper was based on these descriptors.

[2] http://www.rcsb.org

Table 1. CDK2 dataset.

PDB id and Chain	Ligand name	PDB id and Chain	Ligand name	PDB id and Chain	Ligand name	PDB id and Chain	Ligand name
3QL8.A	X01	3QQF.A	X07	3QQG.A	X06	3QQH.A	X0A
3QQJ.A	X11	3QQK.A	X02	3QQL.A	X03	3QRT.A	X14
3QRU.A	X19	3QTQ.A	X35	3QTR.A	X36	3QTS.A	X46
3QTU.A	X44	3QTW.A	X3A	3QTX.A	X43	3QTZ.A	X42
3QU0.A	X40	3QWJ.A	X6A	3QWK.A	X62	3QX2.A	X63
3QX4.A	X4B	3QXO.A	X65	3QXP.A	X64	3QZF.A	X66
3QZG.A	X67	3QZH.A	X69	3QZI.A	X72	3R1Q.A	X75
3R1S.A	X73	3R1Y.A	X76	3R28.A	XA0	3R6X.A	X84
3R71.A	X86	3R73.A	X87	3R7E.A	X88	3R7I.A	X9I
3R7U.A	X96	3R7V.A	Z02	3R7Y.A	Z04	3R83.A	Z14
3R8L.A	Z30	3R8M.A	Z19	3R8P.A	Z46	3R8U.A	Z31
3R8V.A	Z62	3R8Z.A	Z63	3R9D.A	X6B	3R9H.A	Z67
3R9N.A	Z68	3R9O.A	Z71	3RAH.A	O1Z	3RAI.A	X85
3RAK.A	03Z	3RAL.A	04Z	3RJC.A	06Z	3RK5.A	07Z
3RK7.A	08Z	3RK9.A	09Z	3RKB.A	12Z	3RM6.A	18Z
3RM7.A	19Z	3RMF.A	20Z	3RNI.A	21Z	3ROY.A	22Z
3RPO.A	24Z	3RPR.A	25Z	3RPV.A	26Z	3RPY.A	27Z
3RZB.A	02Z	3S00.A	Z60	3S0O.A	50Z	3S1H.A	56Z
3SQQ.A	99Z						

Table 2. Ricin dataset.

PDB id and Chain	Ligand name	PDB id and Chain	Ligand name	PDB id and Chain	Ligand name	PDB id and Chain	Ligand name
1BR5.A	NEO	1BR6.A	PT1	1IFS.A	ADE	1IFU.A	FMC
1IL3.A	7DG	1IL4.A	9DG	1IL5.A	DDP	1IL9.A	MOG
1J1M.A	TRE	1OBT.A	AMP	1RZO.A	GAL	1RZO.B	GAL
1RZO.C	GAL	1RZO.D	GAL	2P8N.A	ADE	3EJ5.X	EJ5
3HIO.A	C2X	3PX8.X	JP2	3PX9.X	JP3	3RTI.A	FMP
3RTI.B	GAL	3RTJ.B	GAL	4ESI.A	0RB	4HUO.X	RS8
4HUP.X	19M	4HV3.A	19L	4HV7.X	19J	4MX1.A	1MX
4MX5.X	5MX						

Table 3. Global network descriptors

Descriptors	CDK2	Ricin
Number of nodes	407	321
Number of edges	459	377
Global clustering coefficient	0.0	0.0
Number of connected components	14	34
Order of largest component	128	134
Order of second largest component	120	28
Order of third largest component	62	18
Density	0.026	0.020
Average clustering	0.748	0.547
Average clustering bipartite graph for ligands	0.823	0.623
Average clustering bipartite graph for proteins	0.188	0.306

2.2 CALI and FSM comparison

In our proposed FSM strategy, we modeled proteins and its ligands as graphs in which atoms are nodes and interactions between atoms are edges. Protein nodes were labeled as positive charged, negative charged, aromatic, hydrophobic, donor or acceptor according to (Sobolev *et al.*, 1999). Ligand nodes were labeled with the same types using PMapper (Pmapper 5.3.8, 2010, Chemaxon[3]) software. Edges (interactions) were labeled according to both a distance criteria (the same adopted in CALI) and the type of its nodes, as aromatic stacking, hydrogen bond, hydrophobic, repulsive and salt bridge.

For each dataset, the resulting graphs were clustered and then we searched for frequent patterns (subgraphs) in each group of graphs using the FSM algorithm gSpan. The clustering analysis is detailed in the paper (Silveira *et al.*, 2015). The CDK2 dataset was segmented in 16 groups while Ricin dataset was segmented in 3 groups.

[3] http://www.chemaxon.com

In transaction based FSM algorithms, including gSpan, given a graph dataset $\gamma = \{G_0, G_1, ..., G_n\}$, $support(g)$ represents the number of graphs in γ which have g as a subgraph. Therefore, this class of algorithm aims to find any subgraph g with $support(g) \geq minSup$ (a minimum support threshold). The support choice is empirical and represents a compromise between the number of patterns and their relevance in FSM algorithms. Increasing the $support$, we have some patterns that are present in more instances from the input dataset. However, among these patters, we have trivial ones (for instance, in our case we have many short patterns - one node - that appear in a large fraction of the input graph dataset). So, we can not just increase the $support$, as we want to find subgraphs that are frequent and also relevant (at least two nodes connected by an edge). So, in our FSM strategy we varied the $support$ value from 0.6 up to 1.0 and summarized results in a table that segments the FSM output by the size of the pattern (subgraph), the $support$ value and the group (from the clustering analysis). This table [4] is coloured and works like a heat map, in which the darker the blue, the higher the $support$. Such table allows a visual inspection and choice of the $support$ value. There are some visual data mining techniques that allow users to perform a exploratory data analysis and to choose appropriate filtering parameters (including support) (Liu *et al.*, 2006; Liu, 2006; DeLine *et al.*, 2015).

Due to this $support$ analysis, in our FSM strategy from (Silveira *et al.*, 2015) the value chosen was $support = 0.7$. The group chosen for discussing FSM results was group 3 in CDK2 dataset and group 2 in Ricin dataset because the protein (PDB id 3QQK) from experimental study for CDK2 (Schonbrunn *et al.*, 2013) was assigned to group 3 in FSM and, similarly, the protein (PDB id 3HIO) from experimental study for Ricin (Ho *et al.*, 2009) was assigned to group 2 in FSM. The same groups were used to compare FSM to CALI results.

3 CALI images

In this work we proposed CALI, a novel graph-based strategy to model protein-ligand interactions and reveal frequent and relevant patterns among them. In addition to the proposed graph based model to summarize and detect common protein-ligand interactions, we also devised a visual interactive representation to illustrate the potential of such modeling. The network visualization is coupled with several filters to support exploration, investigation and analysis of the model and its emerging patterns.

This section provides screenshots that give examples of some features and resources of CALI model and its visual interactive tool. Also, we show some results (residues and atoms important in protein-ligand interaction for CDK2 and Ricin dataset according to experimental studies) that can be obtained by using the filters based in complex network metrics.

Next we present a brief description of each figure.

- Figure 1: visualization of CALI model bipartite graph in which nodes from protein and ligand have different colors and, also, edges have different colors according to the type of interaction.
- Figure 2: example of atom type filtering possibility.
- Figure 3: example of interaction type filtering possibility.
- Figure 4: CALI search example.
- Figure 5: CALI centrality measures filters.
- Figure 6: CALI details obtained on demand.
- Figure 7: Residues from the hinge region of CDK2.
- Figure 8: Important residues in the interaction between Ricin A chain and 28S rRNA.

[4] It is table 6 in Silveira *et al.* (2015) work. Also it is available online at http://homepages.dcc.ufmg.br/~alexandrefassio/biocomp/index.html, choosing *graph pattern table* and then *simple table*.

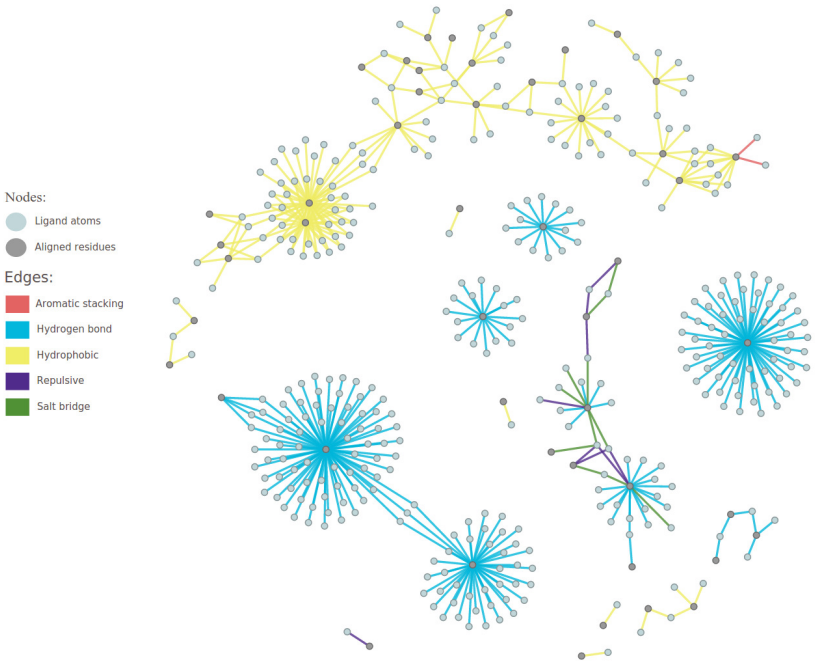

Fig. 1. CALI model biparite graph drawn using a force directed layout. Nodes depict atoms and edges are interactions between them. Different colors distinguish between protein and ligand atoms and also, we have five colors for edges, to represent the five different types of interactions.This graph represents CDK2 dataset.

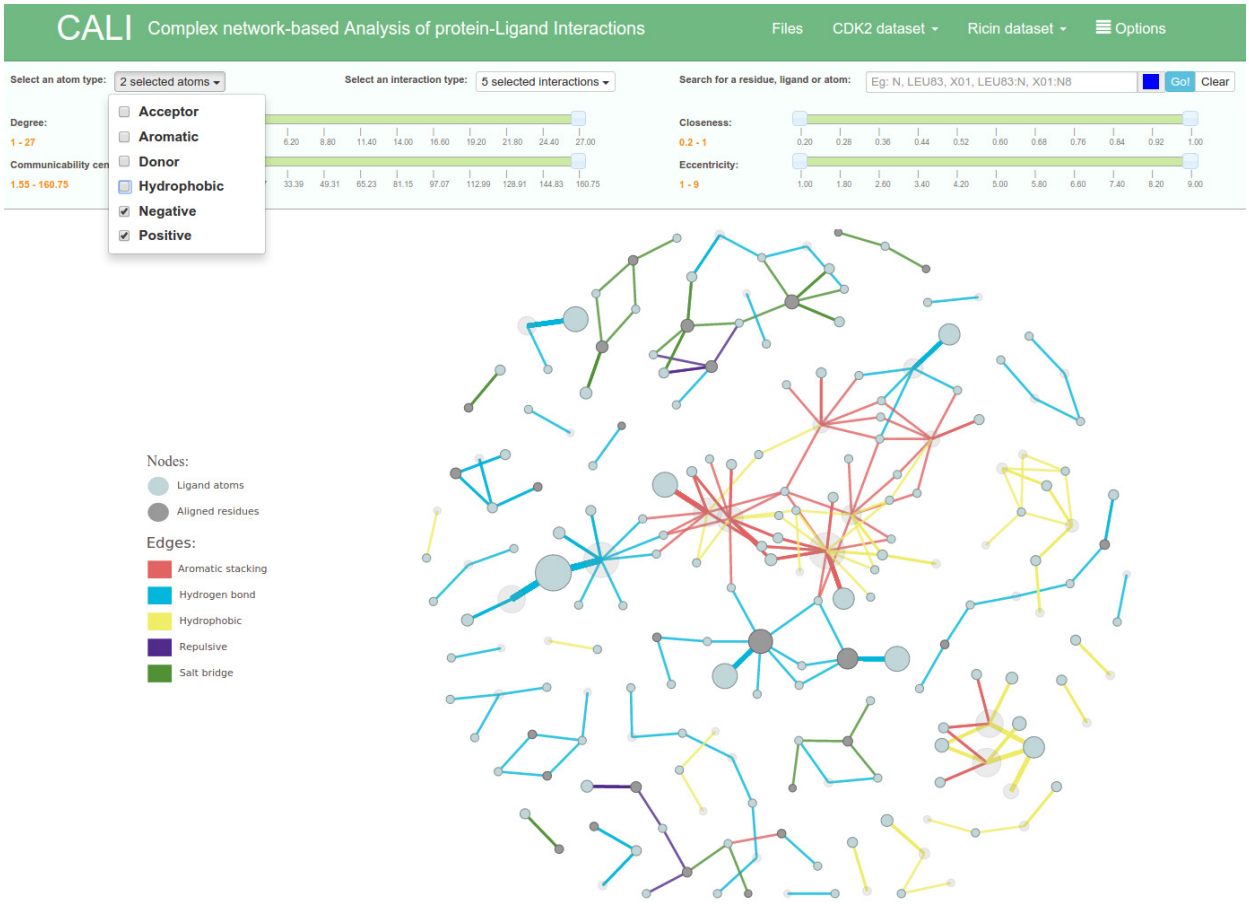

Fig. 2. Example of an atom type filtering possibility. The user can filter out the network by the types of atoms. When he/she checks or unchecks an option, the corresponding atoms (nodes) lose contrast with the background and the others are highlighted. The graph in this example is from Ricin dataset.

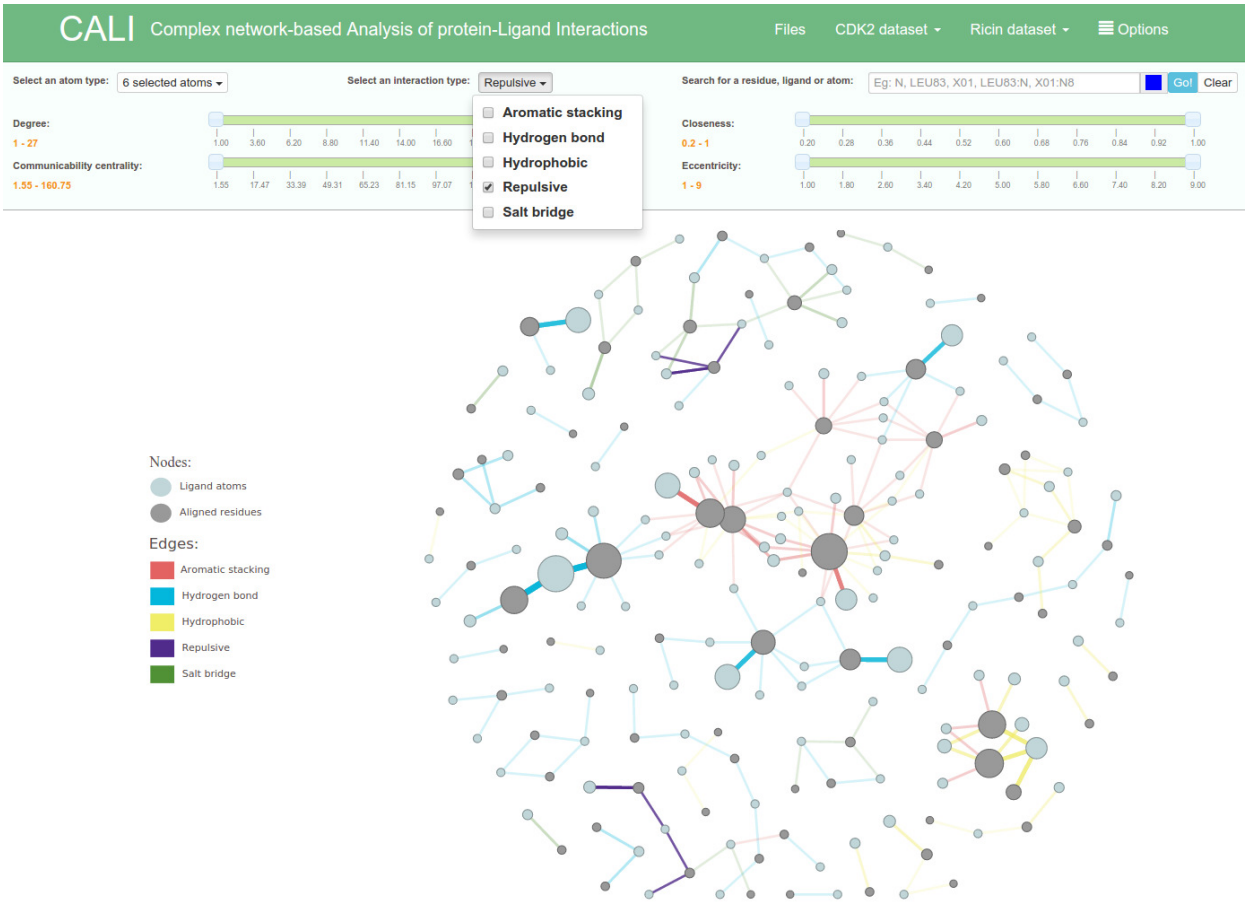

Fig. 3. Example of an interaction type filtering possibility. The user can filter out the network by the types of interactions. When he/she checks or unchecks an option, the corresponding interactions (edges) lose contrast with the background and the others are highlighted. The graph in this example is from Ricin dataset.

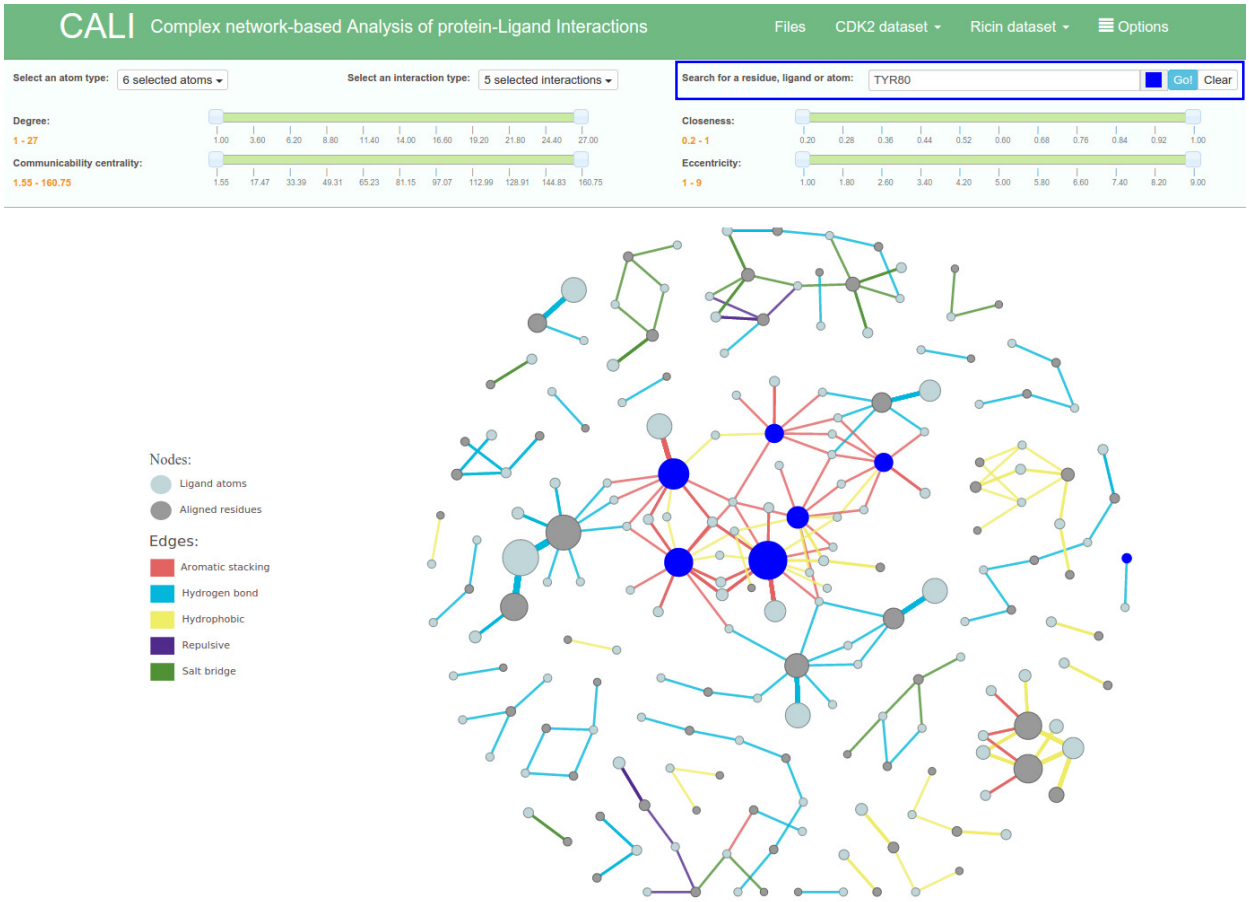

Fig. 4. CALI search example. Users can search for a particular residue and / or atom and it is highlighted (users can pick any color they prefer), which makes easy to find a residue and / or atom in the network. In this figure, we searched for TYR80 and it was highlighted in a brilliant shade of blue. The graph is from Ricin dataset.

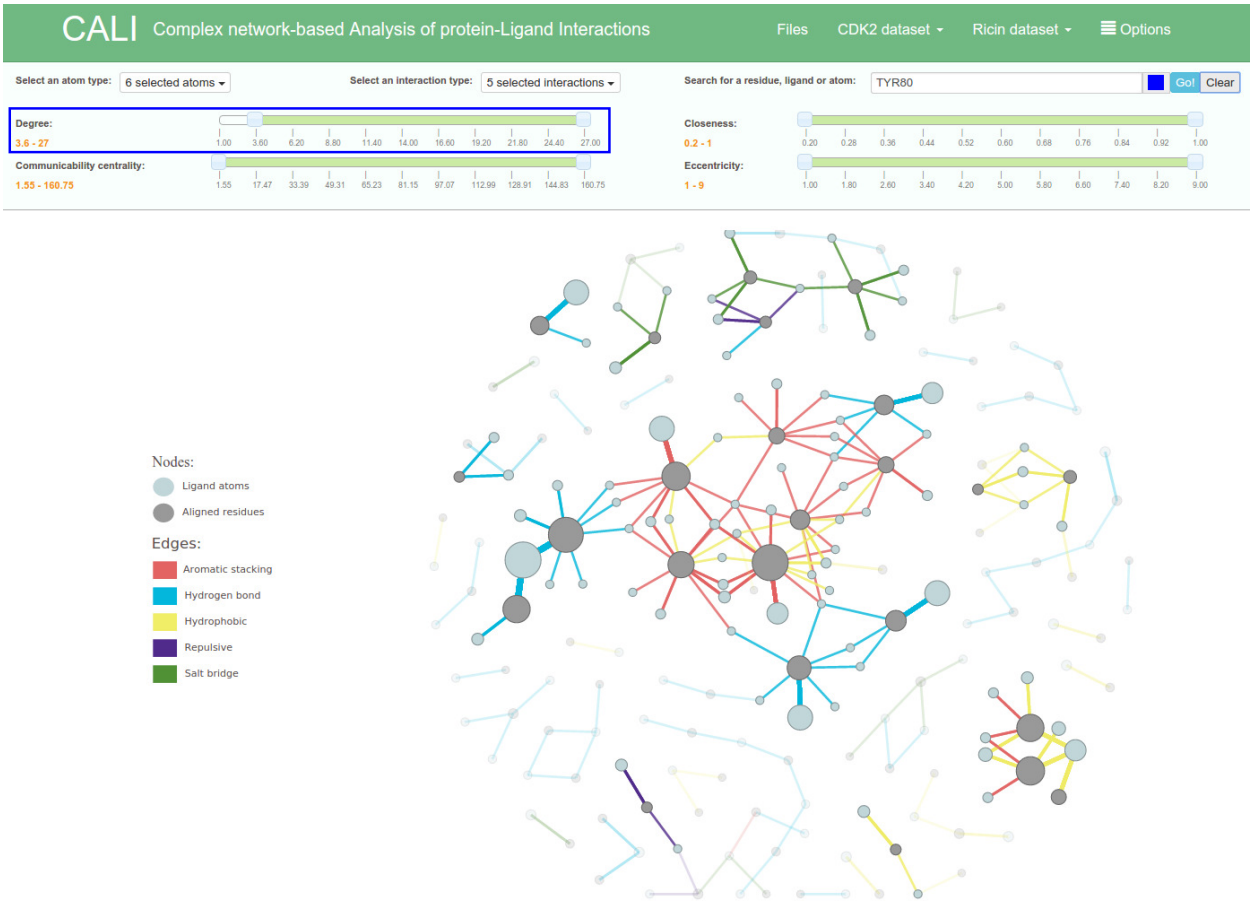

Fig. 5. Centrality measures filters. There are a total of eight different complex network centrality measures that can be used to filter out the network elements through sliders. In this figure, we filter out nodes whose degrees are bellow 10% of the maximum value. The graph is from Ricin dataset.

Fig. 6. Details on demand. For every element of the graph, details can be obtained on demand by passing the mouse over it. In this example, we obtain node details by positioning the mouse over such node. The graph is from Ricin dataset.

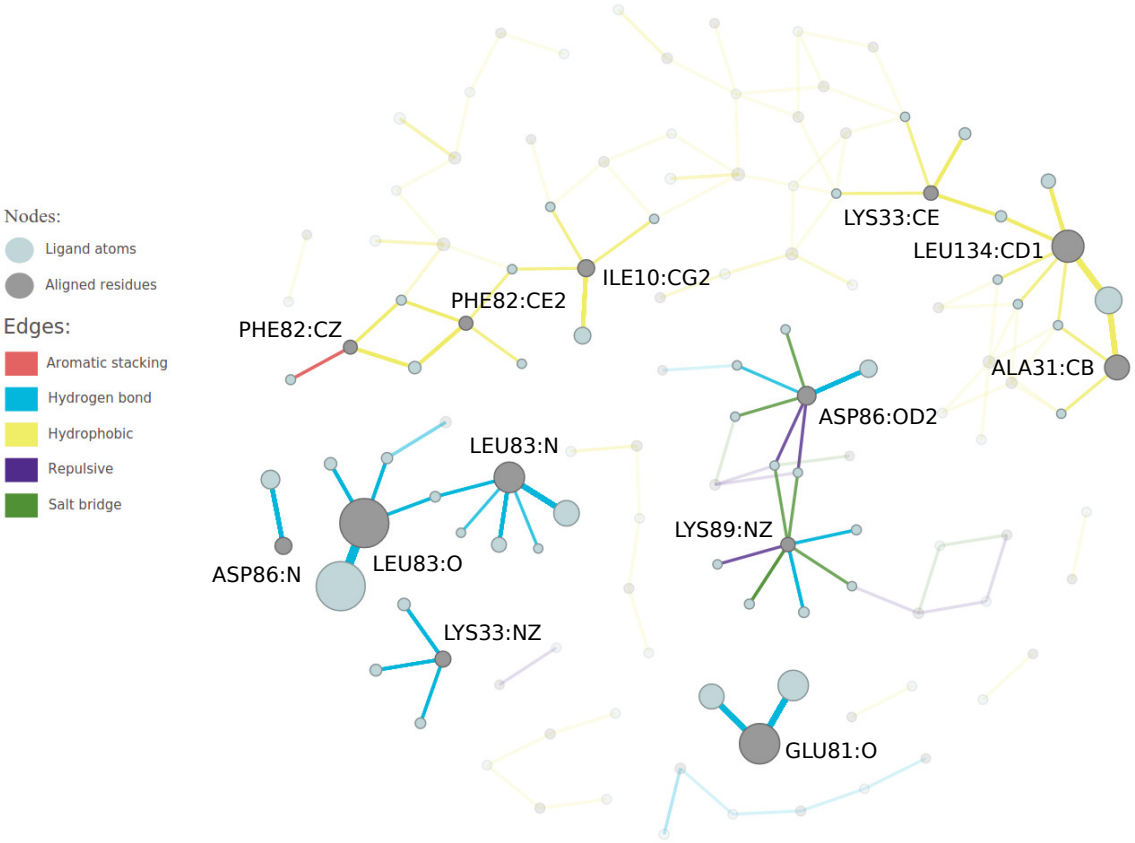

Fig. 7. Residues from the hinge region of CDK2. In this figure, we highlight two important results: (i) CALI was able to spot residues from the hinge region (GLU81, PHE82 and LEU83) according to (Schonbrunn et al., 2013). Moreover, our model was able to spot some residues that frequently interact with ligands through hydrophobic interactions: ILE10, LYS33, ALA31 and LEU134. (ii) In a research done by (Kuhn et al., 2011) of a 3-aminoindazole compound with CDK2 (PDB id 2R64), which is not in our CDK2 dataset, they identified three nitrogen hydrogen bond donors and acceptors that interact with the axis backbone (GLU81 - LEU83). Using CALI, these interactions are easily detected just watching the two components formed in our graph by GLU81 e LEU83. These patterns were obtained by CALI G" model.

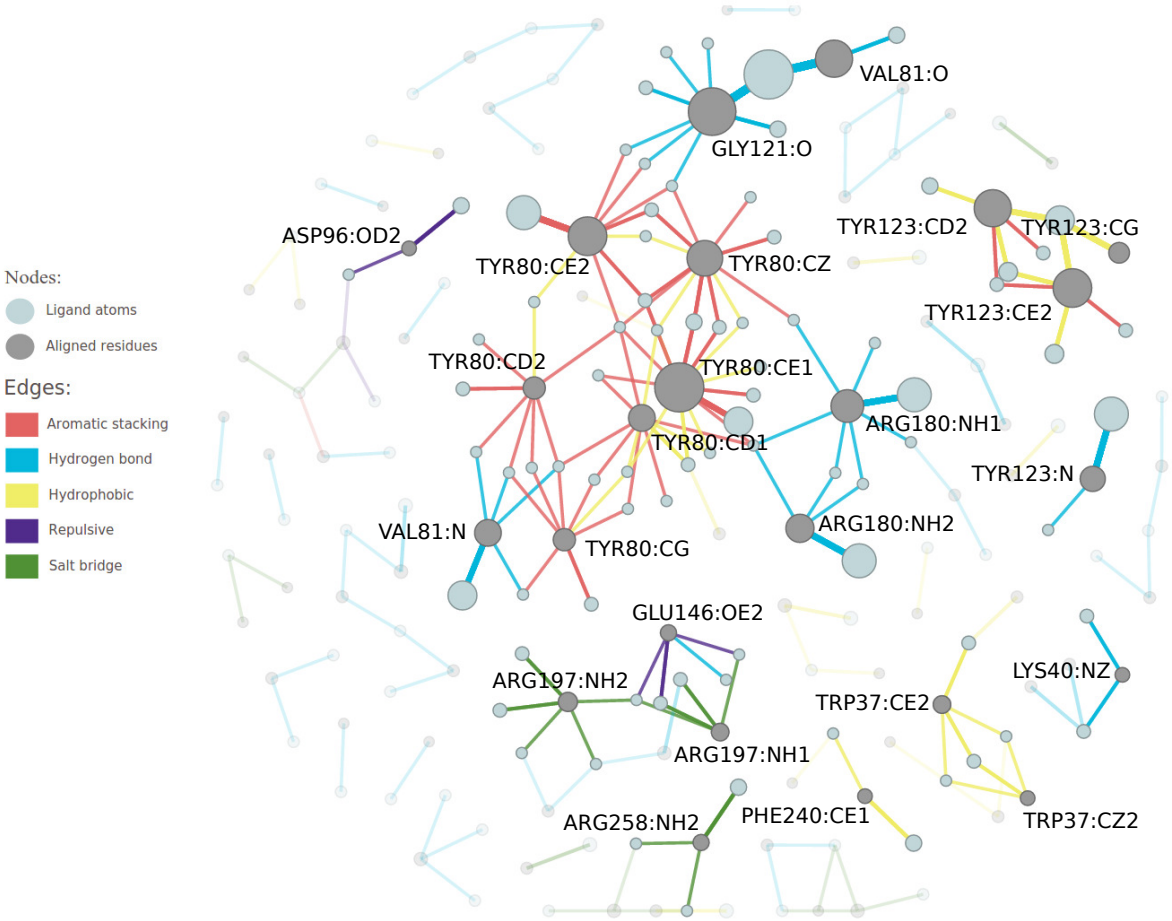

Fig. 8. Important residues in the interaction between Ricin A chain and 28S rRNA. In (Ho et al., 2009), authors co-crystalize RTA with a transition state analogue inhibitor that mimics sarcin-ricin recognition loop of the 28S rRNA. They call our attention to 2 conserved TYR residues (TYR80 and TYR123) establishing π-stacking (aromatic interactions); ARG180 at one end of the π stacking providing cationic polarization and GLU177 serving to activate H_2O nucleophiles. CALI was able to spot the mentioned residues. These patterns were obtained by CALI G" model.

4 Funding

This work has been supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

Bonnici, V. *et al.* (2015). On the variable ordering in subgraph isomorphism algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **13**(9).

Borgelt, C. *et al.* (2002). Mining molecular fragments: Finding relevant substructures of molecules. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 51–58. IEEE.

Cook, D. J. *et al.* (1994). Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, pages 231–255.

DeLine, R. *et al.* (2015). Supporting exploratory data analysis with live programming. In *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*, pages 111–119. IEEE.

Estrada, E. and Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, **77**(3), 036111.

Fischer, I. *et al.* (2004). Graph based molecular data mining-an overview. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 5, pages 4578–4582. IEEE.

Gonçalves, W. R. *et al.* (2015). Pdbest: a user–friendly platform for manipulating and enhancing protein structures. *Bioinformatics*, page btv223.

Ho, M.-C. *et al.* (2009). Transition state analogues in structures of ricin and saporin ribosome-inactivating proteins. *Proceedings of the National Academy of Sciences*, **106**(48), 20276–20281.

Huan, J. *et al.* (2003). Efficient mining of frequent subgraphs in the presence of isomorphism. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 549–552. IEEE.

Huan, J. *et al.* (2004). Spin: mining maximal frequent subgraphs from graph databases. In *Proceedings of the tenth ACM SIGKDD*, pages 581–586. ACM.

Inokuchi, A. *et al.* (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery*, pages 13–23. Springer.

Jiang, C. *et al.* (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, **28**(01), 75–105.

Junker, B. H. *et al.* (2008). *Analysis of Biological Networks*. Wiley.

Koyutürk, M. *et al.* (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, **20**(suppl 1), i200–i207.

Kuhn, B. *et al.* (2011). Rationalizing Tight Ligand Binding through Cooperative Interaction Networks. *Journal of Chemical Information and Modeling*, **51**(12), 3180–3198.

Kuramochi, M. *et al.* (2004). An efficient algorithm for discovering frequent subgraphs. *Knowledge and Data Engineering, IEEE Transactions on*, **16**(9), 1038–1051.

Liu, Y. (2006). *Interactive visual data mining modeling to enhance understanding and effectiveness of the process*. ProQuest.

Liu, Y. *et al.* (2006). Design and evaluation of visualization support to facilitate association rules modeling. *International Journal of Human-Computer Interaction*, **21**(1), 15–38.

Needleman, S. B. *et al.* (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3), 443–453.

Newman, M. (2010). *Networks: An Introduction*. OUP Oxford.

Nijssen, S. *et al.* (2005). The gaston tool for frequent subgraph mining. *Electronic Notes in Theoretical Computer Science*, **127**(1), 77–87.

Pires, D. E. *et al.* (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, **29**(7), 855–861.

Rose, P. W. *et al.* (2015). The rcsb protein data bank: views of structural biology for basic and applied research and education. *Nucleic acids research*, **43**(D1), D345–D356.

Rutenber, E. *et al.* (1991). Crystallographic refinement of ricin to 2.5 å. *Proteins: Structure, Function, and Bioinformatics*, **10**(3), 240–250.

Schonbrunn, E. *et al.* (2013). Development of Highly Potent and Selective Diaminothiazole Inhibitors of Cyclin-Dependent Kinases. *Journal of Medicinal Chemistry*, **56**(10), 3768–3782.

Silveira, S. *et al.* (2015). Revealing protein-ligand interaction patterns through frequent subgraph mining. In *BIOCOMP 2015*.

Sobolev, V. *et al.* (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**(4), 327–332.

Thomas, L. T. *et al.* (2006). Margin: Maximal frequent subgraph mining. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1097–1101. IEEE.

Yan, X. *et al.* (2002). gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 721–724. IEEE.

Yan, X. *et al.* (2003). Closegraph: mining closed frequent graph patterns. In *Proceedings of the ninth ACM SIGKDD*, pages 286–295. ACM.