BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 7 2014, pages 1041–1042
doi:10.1093/bioinformatics/btt741

Data and text mining
Advance Access publication December 20, 2013

SurpriseMe: an integrated tool for network community structure characterization using Surprise maximization

Rodrigo Aldecoa[†] and Ignacio Marín*
Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas (IBV-CSIC), Valencia 46010, Spain

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Detecting communities and densely connected groups may contribute to unravel the underlying relationships among the units present in diverse biological networks (e.g. interactomes, coexpression networks, ecological networks). We recently showed that communities can be precisely characterized by maximizing Surprise, a global network parameter. Here, we present SurpriseMe, a tool that integrates the outputs of seven of the best algorithms available to estimate the maximum Surprise value. SurpriseMe also generates distance matrices that allow visualizing the relationships among the solutions generated by the algorithms. We show that the communities present in small- and medium-sized networks, with up to 10 000 nodes, can be easily characterized: on standard PC computers, these analyses take less than an hour. Also, four of the algorithms may rapidly analyze networks with up to 100 000 nodes, given enough memory resources. Because of its performance and simplicity, SurpriseMe is a reference tool for community structure characterization.

Availability and implementation: SurpriseMe is implemented in Perl and C/C++. It compiles and runs on any UNIX-based operating system, including Linux and Mac OS/X, using standard libraries. The source code is freely and publicly available under the GPL 3.0 license at http://github.com/raldecoa/SurpriseMe/releases.

Contact: imarin@ibv.csic.es

Received on October 11, 2013; revised on December 11, 2013; accepted on December 16, 2013

1 INTRODUCTION

Complex networks are extensively used for representing interactions among elements of a system. This approach is particularly useful in biology; analyzing networks provides relevant information in fields such as genetics (Costanzo *et al*., 2010), ecology (Bascompte *et al*., 2006), neuroscience (Bullmore and Sporns, 2009), systems biology (Barabási and Oltvai, 2004) or proteomics (Schwikowski *et al*., 2000), among others. An interesting property of these networks is the fact that related units of the network tend to create tightly knit groups, usually known as communities. By unraveling the close relationships among certain units, community structure characterization improves our understanding of the system as a whole.

Recently, many strategies have been devised to detect the optimal division of a network into communities. However, none of them alone is able to achieve high-quality solutions in all kinds of networks (Schaub *et al*., 2012; Aldecoa and Marín, 2013a, b). In recent works, we demonstrated that Surprise (S) (Arnau *et al*., 2005; Aldecoa and Marín, 2010, 2011) is an effective measure to evaluate the quality of a partition of a network into communities (Aldecoa and Marín, 2011, 2013a, b). Given such a partition, S calculates the unlikeliness of finding the observed number of intra-community links in a totally random network, according to a cumulative hypergeometric distribution:

$$S = -\log \sum_{j=p}^{min(M,n)} \frac{\binom{M}{j}\binom{F-M}{n-j}}{\binom{F}{n}}$$

Here, F is the maximum possible number of links of the network, n is the actual number of links, M is the maximum possible number of intra-community links and p is the actual number of links within communities. In several complex benchmarks, composed of networks with very different structures, it was shown that the partition of maximum S corresponds to the real community structure, with a minimal/null degree of error (Aldecoa and Marín, 2011, 2013a, b). We also showed that S outperforms modularity (Q; Newman and Girvan, 2004), the most commonly used criterion to define communities, in all benchmarks (Aldecoa and Marín, 2011, 2013a). Although a simple algorithm to maximize S has not yet been devised, we found that choosing among the output of seven high-quality algorithms, the one that provided the maximum S value solved the structure of all networks tested. These algorithms are called CPM (Traag *et al*., 2011), Infomap (Rosvall and Bergstrom, 2008), RB (Reichardt and Bornholdt, 2006), RN (Ronhovde and Nussinov, 2010), RNSC (King *et al*., 2004), SCluster (Aldecoa and Marín, 2010) and UVCluster (Aldecoa and Marín, 2010; Arnau *et al*., 2005). The particular performances of each algorithm are detailed in Aldecoa and Marín (2013a, b).

Here, we present SurpriseMe, a tool integrating those seven algorithms. SurpriseMe accelerates the research process by simply accepting a network as input, internally running all the algorithms and outputting their solutions and their Surprise values. SurpriseMe also calculates distances among the solutions provided by the algorithms, allowing understanding of how congruent they are (Aldecoa and Marín, 2013b).

2 MATERIALS AND METHODS

2.1 SurpriseMe features

SurpriseMe analyses require as an input a text file indicating the list of links that characterize the network. Each line of the file must contain the

*To whom correspondence should be addressed.
[†]Present address: Cooperative Association for Internet Data Analysis, University of California San Diego (CAIDA/UCSD). San Diego, CA. USA.

names of two connected nodes, separated by a tab or space character. From this text file, the software automatically generates the required input formats and runs the algorithms (either all or a subset chosen by the user). Finally, the algorithm that generates the maximum S value is established. For Infomap, RN and RNSC, the unique partitions generated are evaluated, while for RB, CPM, UVCluster and SCluster, which provide several alternative partitions, the ones with the highest S value are considered.

SurpriseMe also compares the solutions of each algorithm, using either the Variation of Information (VI) (Meilă, 2007) or the [1 − NMI] value, where NMI stands for Normalized Mutual Information (Danon *et al*., 2005). In both cases, the greater the value, the more two partitions are different (see Aldecoa and Marín, 2012, 2013a, b). The program also estimates the distances to two additional solutions called *One* (all units of the network are in one community) and *Singles* (each node belongs to a different community). The distances to these two solutions provide further clues of the behavior of the algorithms (Aldecoa and Marín, 2013b). All distances are saved into two distance matrix files (for VI and for 1 − NMI) that can be directly imported into MEGA5 (Tamura *et al*., 2011), a popular free software that allows easily visualizing the hierarchical relationships among the different solutions (see Aldecoa and Marín, 2013b).

2.2 Performance

The algorithms involved are substantially complex. With *n* being the number of nodes and *m* the number of links, here we provide their computational complexities (time/space): for CPM and RB, O(n log n)/O(n + m); for Infomap, O(m)/O(n + m); for RN, O(mˆ1.3)/O(nˆ2); for RNSC, O(nˆ2)/O(n + m); for UVCLUSTER, O(nˆ2 log n)/O(nˆ2); and for SCLUSTER, O(n log n)/O(nˆ2). Therefore, the current version of SurpriseMe is most useful for networks of small to medium size, typically up to 10 000 nodes. We found that very small networks [e.g., those in a typical Girvan and Newman's benchmark (2002), with 128 nodes] are analyzed almost instantaneously. We also tested the software in two more complex standard benchmarks. The first had networks with a Relaxed Caveman (RC) configuration (Watts, 1999) with 10% rewiring, meaning that well-defined communities are present (see Aldecoa and Marín, 2013a). The second was a set of Erdös-Rényi (ER) random graphs (Erdős and Rényi, 1959), essentially without community structure. This last benchmark provides an estimate of the maximum time and resources required. In both cases, we found that networks with up to 10 000 nodes are analyzed by the seven algorithms in less than an hour using a conventional desktop PC, consuming less than 1 GB of memory. However, larger networks require more powerful hardware. With all the programs, a RC network of 50 000 nodes requires about 140 h of analysis and 60 GB of memory. In those cases, it may be advisable to switch off the most time- and resource-consuming programs (RN, SCluster and UVCluster), given that the four remnant programs generally provide very good solutions and are complementary (i.e., they work optimally in different network structures; Aldecoa and Marín 2013a, b). Thus, their combination will still generate either very high or maximum S values. The resources required for a network of 50 000 nodes are then reduced to 40 min and 14 GB of memory (RC structure) or 8 h and 39 GB of memory (ER configuration). For a RC network of 100 000 nodes, we have determined that the four fastest algorithms take 3 h and 30 GB of memory in RC networks and 21 h and 66 GB of memory in ER networks.

3 SUMMARY

Only few researchers have the time and skills to select, download, compile and run multiple community detection algorithms. SurpriseMe allows very simply running a set of state-of-the-art algorithms and determining the one that generates the best Surprise value, i.e. the best partition of the network. It also provides the user with distance matrices (with VI, 1 − NMI values) that help to compare the solutions of the algorithms. Simple to use, it only needs as input a file containing the network to analyze. The well-established power of this type of analysis, together with the simplicity of its use, makes SurpriseMe an excellent tool for characterizing the community structure of complex networks.

Funding: Study supported by grant BFU2011-30063 (Spanish government).

Conflict of Interest: none declared.

REFERENCES

Aldecoa,R. and Marín,I. (2010) Jerarca: Efficient analysis of complex networks using hierarchical clustering. *PLoS One*, **5**, e11585.

Aldecoa,R. and Marín,I. (2011) Deciphering network community structure by surprise. *PLoS One*, **6**, e24195.

Aldecoa,R. and Marín,I. (2012) Closed benchmarks for network community structure characterization. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., **85** (2 Pt. 2), 026109.

Aldecoa,R. and Marín,I. (2013a) Surprise maximization reveals the community structure of complex networks. *Sci. Rep*., **3**, 1060.

Aldecoa,R. and Marín,I. (2013b) Exploring the limits of community detection strategies in complex networks. *Sci. Rep*., **3**, 2216.

Arnau,V. *et al*. (2005) Iterative cluster analysis of protein interaction data. *Bioinformatics*, **21**, 364–378.

Barabási,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*., **5**, 101–113.

Bascompte,J. *et al*. (2006) Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science*, **312**, 431–433.

Bullmore,E. and Sporns,O. (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci*., **10**, 186–198.

Costanzo,M. *et al*. (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.

Danon,L. *et al*. (2005) Comparing community structure identification. *J. Stat. Mech*., P09008.

Erdős,P. and Rényi,A. (1959) On random graphs I. *Publ. Math. Debrecen*, **6**, 290–297.

Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

King,A. *et al*. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**, 3013–3020.

Meilă,M. (2007) Comparing clusterings - an information based distance. *J. Multivariate Anal*., **98**, 873–895.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 026113.

Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., **74**, 016110.

Ronhovde,P. and Nussinov,Z. (2010) Local resolution-limit-free Potts model for community detection. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys*., **81**, 046114.

Rosvall,M. and Bergstrom,C.T. (2008) Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA*, **105**, 1118–1123.

Schaub,M. *et al*. (2012) Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. *PLoS One*, **7**, e32210.

Schwikowski,B. *et al*. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol*., **18**, 1257–1261.

Tamura,K. *et al*. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*., **28**, 2731–2739.

Traag,V.A. *et al*. (2011) Narrow scope for resolution-limit-free community detection. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys*., **84**, 016114.

Watts,D.J. (1999) *Small Worlds: The Dynamics Of Networks Between Order And Randomness*. Princeton University Press, Princeton.