Bioinformatics, 35(24), 2019, 5363–5364
doi: 10.1093/bioinformatics/btz603
Advance Access Publication Date: 2 August 2019
Applications Note

Systems biology

BLANT—fast graphlet sampling tool

Sridevi Maharaj , Brennan Tracy and Wayne B. Hayes*

Department of Computer Science, University of California Irvine, Irvine, CA 92617, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 24, 2019; revised on July 19, 2019; editorial decision on July 29, 2019; accepted on July 30, 2019

Abstract

Summary: BLAST creates local sequence alignments by first building a database of small *k*-letter sub-sequences called *k*-mers. Identical *k*-mers from different regions provide 'seeds' for longer local alignments. This seed-and-extend heuristic makes BLAST extremely fast and has led to its almost exclusive use despite the existence of more accurate, but slower, algorithms. In this paper, we introduce the *Basic Local Alignment for Networks Tool* (BLANT). BLANT is the analog of BLAST, but for networks: given an input graph, it samples small, induced, *k*-node sub-graphs called *k*-graphlets. Graphlets have been used to classify networks, quantify structure, align networks both locally and globally, identify topology-function relationships and build taxonomic trees without the use of sequences. Given an input network, BLANT produces millions of graphlet samples in seconds—orders of magnitude faster than existing methods. BLANT offers sampled graphlets in various forms: distributions of graphlets or their orbits; graphlet degree or graphlet orbit degree vectors, the latter being compatible with ORCA; or an index to be used as the basis for seed-and-extend local alignments. We demonstrate BLANT's usefelness by using its indexing mode to find functional similarity between yeast and human PPI networks.

Availability and implementation: BLANT is written in C and is available at https://github.com/waynebhayes/BLANT/releases.

Contact: whayes@uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A *k-graphlet* is an induced sub-graph g on any set of k connected nodes from a larger graph G, where k has typically been between 2 and 5 (Pržulj *et al.*, 2004). Graphlets have been used to compare and classify networks (Hayes *et al.*, 2013; Yaveroğlu *et al.*, 2014), to identify structure-function relationships (Davis *et al.*, 2015) and for global alignment (Kuchaiev *et al.*, 2010). Supplementary Figure S1 shows all the graphlets on 2, 3, 4 and 5 nodes including their *automorphism orbits* (Pržulj, 2007). Many existing methods that use graphlets for *any* purpose first perform an exhaustive enumeration of all graphlets in the network being analyzed. However, the time complexity for counting all k-graphlets is $O(nd^{k-1})$, where d is the maximum degree in G and n is the number of nodes in G (Shervashidze *et al.*, 2009), this cost is already prohibitive on existing networks. For example, ORCA (Shervashidze *et al.*, 2009) requires 18 h to process the BioGRID human PPI network released in 2018. For many applications, a statistical sample would probably suffice (Chen *et al.*, 2016).

Using a pre-computed lookup table allows graphlet isomorphism to be done in constant time for k-graphlets up to size $k = 8$ (Hasan *et al.*, 2017). Basic local alignment for networks tool (BLANT) leverages this speed, and rather than taking 18 h, it can produce output statistically indistinguishable from ORCA's in minutes. Furthermore, while most other tools only maintain a *count* of graphlets or orbits, BLANT is unique in being able to create a statistically sampled index of nodes belonging to each type of graphlet; this index can form the 'seed' part of seed-and-extend local alignments, or be used to search for structure-function relationships. BLANT has five sampling algorithms with a variety of trade-offs between speed and bias. BLANT provides different output formats: graphlet distributions, sampled indexed graphlet lists, graphlet degree vectors and orbit degree vectors, the latter being compatible with the output format of ORCA. In addition, BLANT supplies a graphlet drawing tool for any k-graphlet, $k \leq 11$.

2 Features

BLANT's command-line interface allows the user to select: (1) graphlet size $3 \leq k \leq 8$ nodes; (2) number of samples; (3) number of threads for additional speedup; (4) graphlet sampling technique; (5) output format; (6) graphlet ID representation and (7) file name.

Four of the sampling techniques implemented in BLANT each begin by selecting an edge uniformly at random to start the graphlet. Then, (1) **Edge Based Expansion (EBE)**: the next edge is selected uniformly at random from edges emanating from the previously selected vertices; this method is fastest on dense networks but produces graphlet distributions with significant bias; (2) **Node Based Expansion (NBE)**: the next node is selected uniformly at random

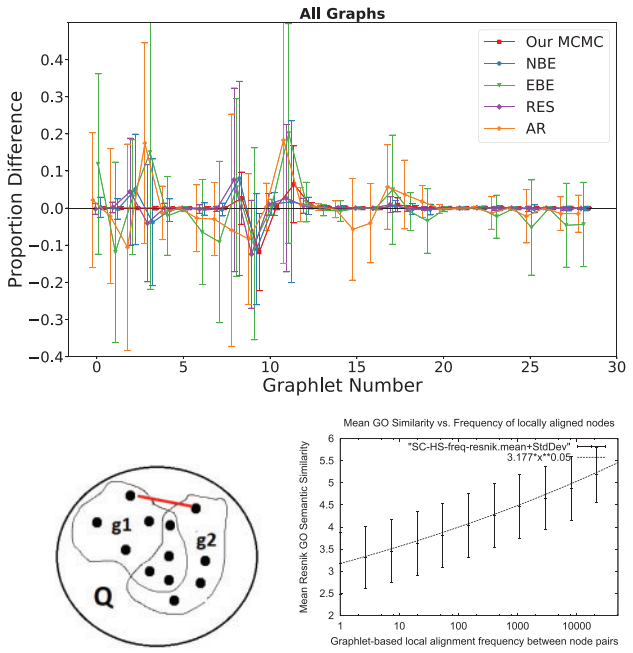

Fig. 1. Top: The error in proportion of each 3, 4 and 5-graphlet obtained from 10^6 graphlet samples using each of BLANT's implemented sampling methods (except AR for which we sampled 10^4 graphlets due to the long run time), on various types of sparse synthetic networks. The error bars represent 1σ of proportion difference. The proportion differences and error bars are smallest for RES and MCMCs, indicating that these methods produce more accurate graphlet concentrations. Our MCMC receives the same results as the original MCMC implementation, indicating correct implementation as shown in Supplementary Figure S2. The sampling error is reported in Supplementary Table S2. **Bottom Left:** Given two k-cliques g_1 and g_2 with overlapping nodes, we can test for the existence of a larger clique Q by examining all remaining connections (e.g. red edge). **Bottom Right:** The number of times a pair of nodes (one in yeast, one in human) are locally aligned in a 6-graphlet orbit is binned on a log scale on the x-axis. Error bars depict the 1σ SD of GO-term similarity for node pairs in each frequency bin. (Color version of this figure is available at *Bioinformatics* online.)

from neighbors of currently selected nodes (Hayes and Maharaj, 2018); this method is fastest on less dense networks and is also less biased than EBE; (3) **Neighbor Reservoir (RES)**: starts with NBE and then further reduces bias by erasing its memory via a random walk for some number of steps. All three of EBE, NBE and RES output a graphlet once k nodes have been found, and then the process starts afresh with a new randomly chosen edge. Thus, all three of EBE, NBE, and RES are equally likely to sample a graphlet from anywhere in G; (4) **MCMC**: a sliding window of $k-1$ edges is kept during a single random walk on the network to form a k-graphlet (Chen *et al.*, 2016); this method produces asymptotically correct graphlet concentrations, but since it is a single long walk, it produces sequences of graphlets in which adjacent graphlets have as many as $k-1$ overlapping nodes and so it may not 'see' the entire graph G unless the walk is extremely long. Finally, the fifth method is impractical but included for completeness: **Accept/Reject** selects k nodes uniformly at random and rejects if the resulting graphlet is not connected (Lu and Bressan, 2012); this method is guaranteed to produce unbiased samples but is extremely slow since the majority of k-node sets are disconnected. Full details of these sampling methods are discussed in Supplementary Section S3. Supplementary Table S1 shows the time taken to sample differently sized graphlets from various networks using each sampling method. Figure 1 (top) shows the difference in proportion from the true proportion (computed by ORCA), of each graphlet obtained by using each of these five sampling techniques on various synthetic networks described in Supplementary Section S5.

The sampled graphlets may be output in various formats: (1) indexed k-graphlet lists: each line contains $k+1$ columns; the first column contains the graphlet IDs and the next k columns are the vertices forming the graphlet, (2) k-graphlet counts: the total of each type of k-graphlet sampled from the network or the concentration of each k-graphlet in the case of MCMC sampling, (3) sampled graphlet degree vectors: a vector for each node representing the number of sampled graphlets to which it belongs, (4) sampled orbit degree vectors: a vector for each node representing the number of orbits it touches from the sample. BLANT's setup instructions are in Supplementary Section S4, its options and interface are shown in Supplementary Figure S4.

3 Biological relevance

GO Term Prediction: In a PPI network, a large clique suggests a protein complex whose members share common function. Finding large cliques is NP-complete. By sampling 10^7 8-node graphlets from the 2017 BioGRID human network, and analyzing overlapping 8-node cliques (see bottom left of Fig. 1), we found a 60-node near-clique (having 97% of all possible edges). Using GO terms (The Gene Ontology Consortium, 2008) of 2016, and assuming any GO term appearing in more than half the 60 nodes should be transferred to the rest, resulted in 213 GO term predictions, 46 of which were corroborated by GO terms of 2018. None of the remaining predictions were contradicted, suggesting they may be corroborated by future GO discoveries.

Topology-function relations: We indexed 10^6 6-node graphlets from both the 2018 BioGRID yeast and human PPI networks. Whenever the same graphlet appeared in both networks, we imposed the resulting local alignment between them. If function is related to topology, then we expect a pair of frequently aligned nodes (one from yeast, one from human) to share functional similarity. Figure 1 (bottom right) shows that mean Resnik similarity increases with pairwise alignment frequency. To our knowledge, this is the first demonstration of such a broad correlation between local network topology and functional similarity.

Conflict of Interest: none declared.

References

Chen,X. *et al.* (2016) A general framework for estimating graphlet statistics via random walk. *Proc. VLDB Endowment*, **10**, 253–264.

Davis,D. *et al.* (2015) Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, **31**, 1632–1639.

Hasan,A. *et al.* (2017) Graphettes: constant-time determination of graphlet and orbit identity including (possibly disconnected) graphlets up to size 8. *PLoS One*, **12**, e0181570.

Hayes,W. and Maharaj,S. (2018) BLANT: sampling graphlets in a flash. In: *q-bio, Rice University, Houston, Texas, USA*.

Hayes,W. *et al.* (2013) Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, **29**, 483–491.

Kuchaiev,O. *et al.* (2010) Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interf.*, **7**, 1341–1354.

Lu,X. and Bressan,S. (2012) Sampling connected induced subgraphs uniformly at random. In: *International Conference on Scientific and Statistical Database Management*. Springer, Berlin, Heidelberg, pp. 195–212.

Pržulj,N. *et al.* (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

Pržulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.

Shervashidze,N. *et al.* (2009) Efficient graphlet kernels for large graph comparison. In: *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA*. Volume 5 of JMLR: W&CP 5, pp. 488–495.

The Gene Ontology Consortium. (2008) The gene ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Yaveroğlu,N. *et al.* (2014) Revealing the hidden language of complex networks. *Sci. Rep.*, **4**, 4547.