Poster J-25

Biological Network Comparison Using Graphlet Degree Distribution

Authors:
Natasa Przulj (*UC Irvine, Computer Science*)

Short Abstract: We devise a highly constraining network comparison metric based on local structural properties that are a direct generalization of the degree distribution. We use this new metric to demonstrate that geometric random graphs better model protein-protein interaction networks of various species than do Erdos-Renyi, random scale-free, or Barabasi-Albert scale-free networks.

Long Abstract:
We devise a highly constraining network comparison metric based on local structural properties that are a direct generalization of the degree distribution. We use this new metric to demonstrate that geometric random graphs (GEO) better model protein-protein interaction (PPI) networks than do Erds-Renyi (ER), random scale-free (RSF), or Barabasi-Albert scale-free networks (BASF).

The full-scale comparison of two arbitrary networks is computationally intractable, because it contains the subgraph isomorphism problem, which is NP-complete. Thus, heuristic measures must be developed for network comparison. Our metric, based on the graphlet degree distribution (GDD), measures a network's local structure, which is appropriate for networks in which various local areas have been intensely mapped but for which the global structure is unknown. Recall that the traditional degree distribution measures the number of nodes of degree k for each value of k; in other words, it measures the number of nodes that are ``touching'' k edges, for each value of k. Note that an edge is the smallest non-trivial example of a connected graph. We refer to small connected graphs as graphlets. More specifically, graphlets are small connected non-isomorphic induced subgraphs of a large network (Przulj et al. 2004). We denote the 2-, 3-, 4-, and 5-node graphlets by G_0, G_1,…, G_29. The traditional degree distribution measures how many nodes ``touch'' k G_0s (i.e, edges), for each value of k. Applying a similar measurement to the twenty-nine 3-, 4-, and 5-node graphlets G_1, G_2,… G_29 gives us a set of distributions, which we call the graphlet degree distributions. Note, however, that there are more than 29 distributions, because, for example, it is topologically relevant to distinguish between nodes touching a 3-node path at an end, or at the middle. This is due to the 3-node path admitting an automorphism that maps its end nodes to each other and the middle node to itself. An automorphism g is a map from a graph onto itself that permutes the nodes while preserving edges; i.e., an edge xy is in the original graph if and only if the edge g(x)g(y) is in the automorph. When a set of nodes can be mapped to each other under an automorphism group, we say that those nodes belong to a single automorphism orbit. For example, end nodes of a 3-node path belong to one orbit, while the mid-node of it belongs to another. Graphlet G_0 (the edge) has only one orbit, as does graphlet G_2 (the triangle); graphlets G_1 (the 3-node path), G_3 (the 4-node path) and G_4 (the “claw”) have two orbits, etc. Analogous to the degree distribution, for each of the 73 orbits of 2-, 3-, 4-, and 5-node

graphlets, we count the number of nodes touching a particular graphlet at a particular orbit. Thus, the traditional degree distribution is the first one in the spectrum of 73 ``graphlet degree distributions (GDDs)''.

To compare two networks, we first compute all 73 GDDs of each. We normalize the 2 x 73 distributions each to have a total area of unity, to meaningfully compare distributions with different areas. We compute a_j, the agreement in distribution j, for j=0,…,72, where a_j=1 means that the networks have identical normalized distributions, and a_j=0 means the networks are very different. Finally, the average agreement across all a_j provides one number that compares two networks (Przulj 2006, submitted).

We computed GDDs of 14 PPI networks belonging to 4 different organisms (yeast, fruitfly, worm, and human), obtained by various high-throughput experiments (Y2H, TAP, HMS-PCI) and by human curation (BIND, HPRD, MINT), and compared them with 1,400 model networks belonging to 4 different random graph models (Erdos-Renyi, random and Barabasi-Albert scale-free, and geometric random). Out of the 14
PPI networks, geometric random graph models had significantly better agreement with the data than all the other models in 12 cases, and had comparable agreement with the best model in the other 2 cases.

We anticipate that network comparison will soon equal sequence comparison in importance for enhancing biological understanding.

References:
N. Przulj, D. G. Corneil, I. Jurisica. ``Modeling interactome: Scale-free or geometric?'' Bioinformatics, 20(18):3508-3515, 2004.
N. Przulj. ``Biological Network Comparison Using Graphlet Degree Distribution'', submitted, 2006.

Acknowledgement:
We thank Derek Corneil and Wayne Hayes for helpful comments and discussions, Pierre Baldi for providing computing resources, and Jason Lai for help with programming.