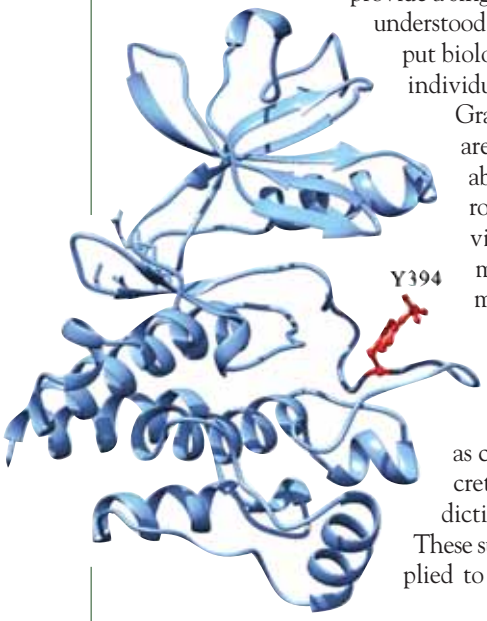
Under TheHood

BY JOSE LUGO-MARTINEZ AND PREDRAG RADIVOJAC, PhD

Vertex Classification in Graphs

Graphs, or networks, have been widely adopted in computational biology, with examples including protein-protein interaction networks, gene regulatory networks, and residue interaction networks in proteins, to name a few. Graphs provide a single and methodologically well-understood way to describe high-throughput biological data as well as data from individual experiments.



A structure of lymphocyte-specific protein tyrosine kinase (PDB id: 3lck) with a highlighted residue (Y394) that is known to be an autophosphorylation site.

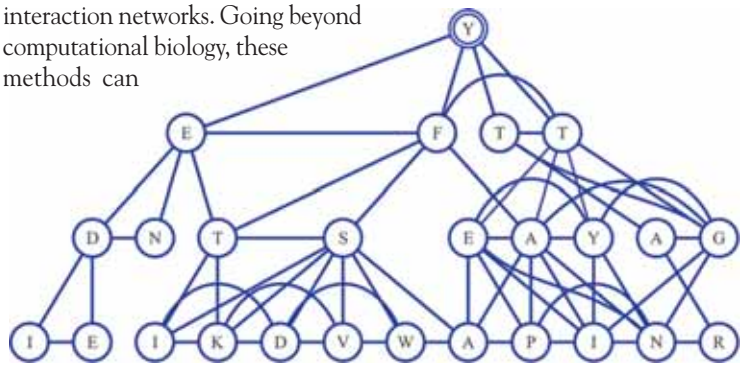
Graphs are most useful when they are analyzed to draw inferences about the data. Such analyses fall roughly into two camps: unsupervised techniques for network motif finding (graphs that occur more frequently than expected) and clustering (grouping of data); and supervised techniques, which usually involve prediction tasks such as classification (prediction of discrete outputs) and regression (prediction of continuous outputs). These supervised techniques can be applied to predict properties of a graph (graph classification) or of the vertices in a single graph (vertex classification). Below we describe how vertex classification techniques can be used to gain new insights into the residues that make up a protein.

When using graphs to analyze protein structures, the first step is to convert each protein structure of interest into a residue interaction network, where vertices represent amino acid residues and the links between pairs of vertices indicate that the two residues are in contact—often if the distance between them is within 3 to 6Å.

In the graph classification scenario, each protein can be seen as a different graph and the task may be to predict a structural or functional classification of such a protein, or graph—e.g., its fold class (e.g., barrel, globin) or its cellular role (e.g., catalytic activity, transcription factor activity). On the other hand, in the vertex classification scenario, all proteins are collectively considered as a single large disconnected graph, and the objective may be to predict some properties of interest regarding each residue. For example, the identification of functional residues (e.g., DNA-binding residues, post-translationally modified sites, etc.) falls under the vertex classification scenario, an example of which is shown here.

There are three principal approaches to vertex classification. First, for properties that tend to be localized, probabilistic graphical models (e.g., Markov Random Fields) can be used to propagate class labels across a graph, for example, from a group of DNA-binding residues to their neighboring vertices. Second, one can map each vertex together with its local neighborhood into a vector in the Euclidean space and then use standard machine-learning techniques for predictor development. Here, vertex properties such as degree, clustering coefficient, and others might be used to encode each vertex into a fixed-dimensional vector. Third and finally, if one has insight into how to effectively measure similarity between vertex neighborhoods, one might define a kernel (similarity) function over pairs of vertices based on their graph neighborhoods, for example, one based on simultaneous random walks starting at the two vertices of interest. Kernel functions can then be used by learning algorithms capable of working with similarities between objects rather than sets of object descriptors. In contrast to probabilistic graphical models, the latter two approaches assign class labels based on the similarity of vertex neighborhoods regardless of their location in the graph; however, they may be less effective in modeling dependencies between vertices. The final choice of a method thus depends on the problem at hand and domain knowledge.

In addition to identifying functional residues in protein structures, vertex classification is helpful for predicting protein function or disease associations from protein-protein interaction networks. Going beyond computational biology, these methods can also help identify malicious web sites on the Internet or predict a person's voting preferences in a social network. As the volume and nature of data change with technology, development of vertex classification methods that can handle real-life (big and noisy) data, incorporate the wealth of auxiliary domain information in principled ways, and/or increase the efficiency of learning and inference will have wide implications not only for computational biology, but also for a number of scientific and industrial applications. □



The local graph neighborhood for the Y394 residue (double circled). The graph was generated using a distance threshold of 6Å. Each residue is represented by a single letter amino acid code but the positional information is removed. The task of a classifier is to predict class labels (here, presence or absence of phosphorylation) for each vertex (local graph neighborhood) in the residue interaction network.

DETAILS

Jose Lugo-Martinez is a PhD candidate in computer science and Predrag Radivojac is associate professor of computer science and informatics at Indiana University, Bloomington.