Results:

Database description:

We combined data from a The Cancer Genome Atlas (TCGA) pan-cancer investigation of 9 423 tumour exomes to uncover novel cancer driver genes. The network represents our understanding of gene-protein interactions. The network has an average degree of 61.02 (128.33), and Figure 2a shows the sizes of the k-hop neighbourhoods. Each node in the network is represented by its MutSig P-value, which has been log10 converted.These P-values indicate if a gene's mutational patterns in tumours and normal tissues differ significantly. We have access to P-values for 18 154 genes in total. We eliminate any nodes from the network that cannot be represented with a MutSig P-value, as well as all isolated nodes, as a pre-processing step. As a result, there are 11 449 genes in the InBio Map network that have been tested with the MutSig tool and are related to at least one additional node. These are our cancer driver gene candidates for network-based prediction.

Class labels:

To train a classifier, supervised machine learning often requires access to labelled data. The CGC data from the COSMIC database is used to obtain gene labels. We used this set as our ground truth and retrieved a list of 723 genes that have been causally linked to cancer. Genes in the CGC are divided into Tiers 1 and 2, with Tier 1 showing confirmed cancer-related activity and Tier 2 showing strong indications of playing a role in cancer. Both tiers are treated similarly in our analysis. Our network overlaps with the set of 723 genes, yielding a total of 635 cancer genes. As a result, our dataset contains approximately 6.0 percent ‘positive' samples. The remaining genes are known as unlabeled genes, and we're looking for new cancer genes among them.

Experimental setup:

We created a cross-validation approach based on repeated undersampling of the majority class to solve the above-mentioned issues provided by the class imbalance and lack of a negative class. Figure 3 shows the cross-validation technique, and Supplementary Algorithm S1 contains a pseudocode for it. The dataset is represented as a DR11 449d matrix, where d is the number of node characteristics.

Discussion and conclusions:

By combining MutSig summary statistics  with PPI networks, we provided a unique methodology for identifying cancer driver genes in this research . We devised a unique node embedding strategy (MoPro embeddings) to enable supervised categorization of cancer driver genes, in contrast to state-of-the-art methodologies that set out to solve this problem with unsupervised procedures. Incorporating knowledge of the data distributions of well-established cancer driver genes into the issue of cancer-gene prediction in a supervised manner allows us to learn from what we already know: we include knowledge of the data distributions of well-established cancer driver genes and learn from these distributions to better the prediction task.