Application of Artificial Intelligence to Predict a New Class of Novel Synthetic Lethal Targets

Spyro Mousses[1], David Schneider[2], **Jeff Kiefer[1]**, Pieter Derdeyn[2], Kendyl Douglas[2], Abhishek Kothari[2], Daniel D. Von Hoff[3], Chris Yoo[2]
[1]Systems Oncology, Scottsdale, AZ; [2]Systems Imagination, Inc., Tempe, AZ; [3]Translational Genomics Research Institute, Phoenix, AZ.

Abstract

Background: Synthetic lethal targets are proteins that are contextually vulnerable. Inhibitors of PARP1, for example, selectively produce a lethal phenotype in the context of cancer cells which have lost BRCA1 or BRCA2 function. As a high mutation rate is a hallmark of many cancers, targeting synthetic lethal interactions to selectively inhibit cancer cells with altered genetic backgrounds may increase the specificity and efficacy of therapeutics. Recently, clinical trials have targeted synthetic lethal pairs such as EGFR and BRAF, TP53 and BCL2, and PTEN and CHD1. Previous attempts to identify synthetic lethal targets have relied on empirical results from published studies of biological pathways perturbed in cancer cells. Developing strategies to rapidly identify synthetic lethal pairs by combining multiple experimental and computational approaches would result in a new class of potential cancer drug targets beyond the existing efforts that rely on single experimental or computational methods alone. **Methods**: Here we present Expansive AI, an artificial intelligence augmented knowledge network that enables rapid hypothesis generation for accelerated discovery research. Using a purpose-built hypergraph database of massive, integrated genomic and biomedical data, we can query all synthetic lethal pairs and their component genes, as well as a wealth of data related to these genes. The database of biological data includes 11,000+ cancer genomes from TCGA, prior knowledge resources such as gene ontology and pathway resources, and experimental data including chemical and protein interaction and patent data. The hypergraph's architecture allows for linking and nesting data, enabling efficient extraction of biologically-relevant features. **Results:** Using these features, a neural network classified 798 new candidate pairs that have previously not been reported. The candidate pairs were filtered to include only known tumor suppressor and amplified genes. This produced a list of gene pairs which may represent the most novel class of synthetic lethal target candidates identified to date. **Conclusions:** We highlight the results of this AI-based approach and discuss validation efforts of the predicted interactions in specific cancer contexts.

Materials

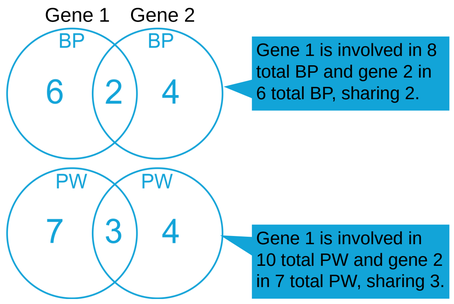
- Hypergraph database (Hyperion v.2.34) consisting of 5,277,455 hyperedges with one entity (e.g. "nodes") and 3,157,032 hyperedges with 2 or more entities, linking data from over a dozen cancer-specific knowledge resources (e.g., PubChem, Pharos, TCGA, etc.).
- NVIDIA DGX Workstation (4X Tesla V100 GPUs, 128 GB GPU RAM, 2,560 Tensor cores, 20,480 CUDA cores, Intel Xeon E5 2.2 GHz 20-core, 256 GB system RAM, running Ubuntu Linux OS 16.04.4 LTS)
- The hypergraph is indexed along every hyperedge across every combination of attributes for each hyperedge. As such, any search on any attribute results in a parallel search with parallel merge-sort. This approach permits high speed results for even large queries. The query used for this effort was "all protein coding gene pairs" - this resulted in all the data in the hypergraph for every possible protein coding gene pair (~440 million pairs).
- Synthetic lethal datasets curated from SynLethDB (http://bit.ly/2YVJHXT) and SLOrth (http://bit.ly/2HHCBk8), were integrated into the Hypergraph.

Neural Network

- Multi-GPU multi-class multi-label classifier used.
- Software used: Keras, Pandas, scikit-learn, MxNet.
- The high level open-source software package Keras was used because of efficient use of multi-GPUs on DGX platform for multi-class classification
- Open source deep learning framework, MxNet used with Module and Gluon APIs.
- Training performed on random sample of knowns plus balanced set of unknowns.
- Softmax cross entropy loss used in prediction of probability of each output class.
- 18 neuron input later, 3 hidden layers (40, 50, and 60 neurons), 6 output neurons (6 classes).
- 6 classes (1 = known, 2-6 prediction of lower quality by class).

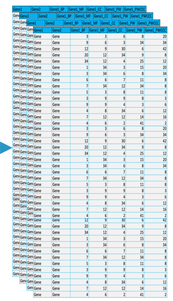
Data Preparation

- Extract script written to retrieve every protein coding gene (~22,000), and related molecular entity type:
  - Gene Ontology (Biological Processes (BP), Molecular Function (MF), Cellular Component (CC)), Pathway (PW), Pathway-Cellular Component (PWCC), Protein Complexes (CMPLX).
- Count number of related molecular entities by type:
  - So if a given gene is related to 4 distinct BPs, the count for that gene for entity-type BP is 4.
- For each gene pair (~440 million), count the number of related molecular entities by entity type that both genes are involved in. Multiply count by 10.
  - So if 2 genes in a gene pair are both involved in the same 2 BP, the entity-type score for BP for the gene pair is 20 (2x10).
- In summary, for every gene-pair there are 18 features–each of which correlates to an entity-type. 6 (BP, MF, CC, PW, PWCC, CMPLX) for gene 1, 6 for gene 2, and 6 for both genes.



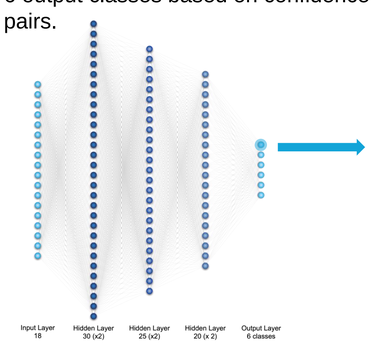
Gene1	Gene2	Gene1_BP	Gene2_BP	Gene1_PW	Gene2_PW	Pair_PW	Pair_BP
Gene1	Gene2	8	6	10	7	30	20

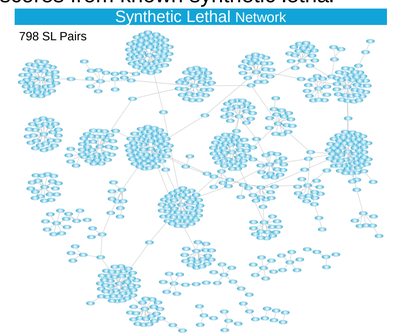
Data to Load in Neural Network



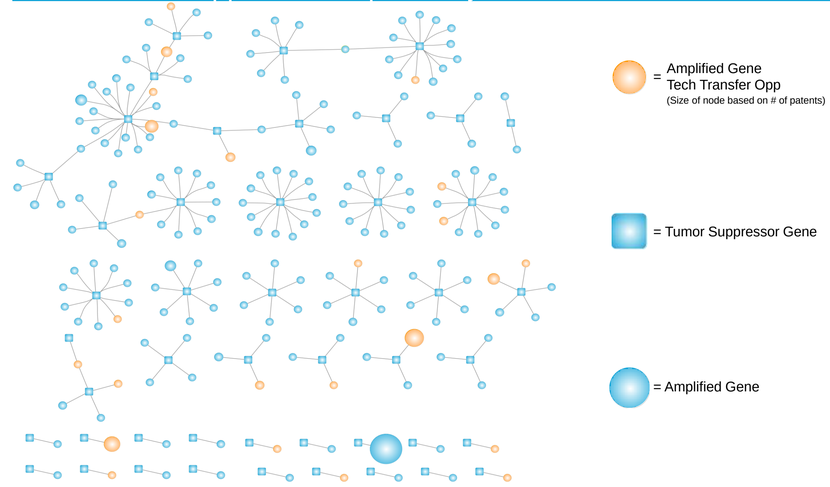
Neural Network Architecture

- 18 input nodes for all protein coding genes and gene-gene pairs.
- Neural network trained on known synthetic lethal data in public domain.
- 6 output classes based on confidence scores from known synthetic lethal pairs.





Tumor Suppressor Amplified Synthetic Lethal Network



Intellectual Property & Licensing Opportunities

US Technology Transfer Opportunities	Target Cancer Treatment Patents (Priority 2012, >5)
24	46

Amplified Gene Tractability

Small Molecule Clinical Precedence	Small Molecule Predicted Tractability	Monoclonal Antibody Clinical Precedence	Monoclonal Antibody Predicted Tractability
21	68	5	66

Conclusions

- Hypergraph database for integrated storage of biomedical data allows for efficient storage, analysis, and retrieval for sophisticated analysis.
- NVIDIA DGX GPU dependent workstation increases computation speed at reduced cost.
- Neural network-based classifier identified 798 novel SL pairs.
- Filtering of SL network to TSG-AMP pairs prioritize novel drug targets.