Determining the Optimal Embedding Technique for Mapping Gene Expression Samples into a Distributed Ontology Space

There are more than 1 million publicly available human gene expression samples. These samples together constitute an extremely valuable resource that any researcher can use to gain new biological insights about genes and cellular mechanisms. However, most of these samples lack systematic annotation of the context in which they were generated including, tissue and cell type of origin as well as disease state. Developing computational methods that can automatically label a sample is a grand challenge in biomedical research.

Previous methods have shown that building models that utilize gene expression profiles from similarly labeled profiles is a powerful way to auto generate labels. However, most of these previous methods do not intrinsically take into account the fact that tissues, cell types or diseases are related to each other in terms of anatomy, physiology or manifestation. Also, based on the way these methods are designed, one cannot make predictions for tissues, cell types or diseases that did not have a substantial number of manually annotated samples to begin with.

Recently, a new method (termed *OnClass*[1]) has been developed to predict the labels of single cell gene expression profiles. Typically, sample classification methods build models that directly annotate a given gene expression profile to a predetermined set of labels. The OnClass method instead uses a neural network to first generate a mapping between gene expression space to a label space that is created through a low dimensional embedding of an ontology. This mapping simultaneously captures the relationships of different terms in the ontology while also allowing predictions for terms that were never seen during training the model.

While the *OnClass* method has shown state-of-the-art results, there was little investigation into how best to embed an ontology that optimizes the mapping from gene expression space. To answer this question, we systematically evaluated different embedding techniques and embedding dimension sizes **[Fig. 1]**. The embedding techniques fall in one of two categories: walk-based embeddings specifically designed for graphs and more traditional dimensionality reduction methods. In general, walk-based embeddings rely on the *word2vec* method, where first a set of random walks are generated (sentences in *word2vec*) that are used to determine positive node term-term pairs (word pairs in *word2vec*) to train a neural network which generates a low dimensional embedding. These methods can use more standard random walks (*deepwalk*), allow for choosing between breadth- or depth-first searching (*node2vec*), or take into account sub-structure within the graph (*struct2vec*). Regardless of the exact technique, these methods all generate a low dimension vector for every node that preserves the structure of the underlying graph. In addition, we evaluate traditional dimensionality reduction methods. Here, we either represent the ontology as an adjacency matrix and then directly apply *PCA* or *SVD*, or perform *PCA* or *SVD* on a diffusion matrix that is generated from the adjacency matrix via a random-walk with restart kernel (*PCAdiffusion* or *SVDdiffusion*).

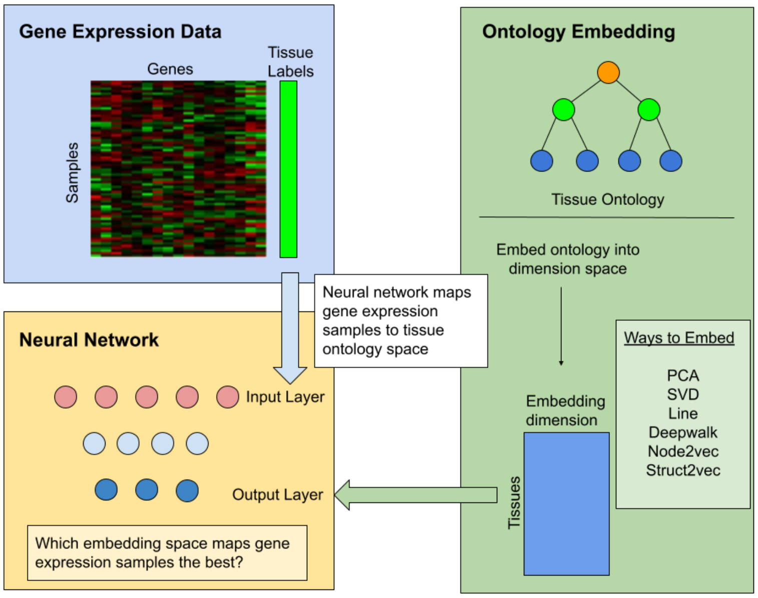

Figure 1. Workflow for systematic evaluation of ontology embedding techniques.

For evaluating the effectiveness of these embedding techniques, we used data from the Affymetrix Human Genome U133 Plus 2.0 Array, consisting of 8,416 gene expression samples with known tissue labels from 52 different tissues. The samples were split into training, validation, and testing sets, with the splitting done in such a manner that all expression profiles (GSMs) from the same experiment (GSE) were placed into the same split. To test the prediction on tissues unseen in the training data, 16 tissues were exclusively placed in the testing split. The specific set of genes used as features in the neural network are the LINCS "landmark" genes.

The neural network used to train each model was implemented using the Python package Keras. Each model contained a single hidden layer, used the Adam optimizer with a learning rate of 1e-05, and a customized loss function. This loss function first finds the similarity of the output vector that is generated for a given expression sample to the embedding vector for each tissue in the label set. Subsequently, a prediction is generated using categorical cross entropy. In addition to testing different ontology embedding techniques, we also systematically tuned both the size of the hidden layer as well as the embedding dimension size.

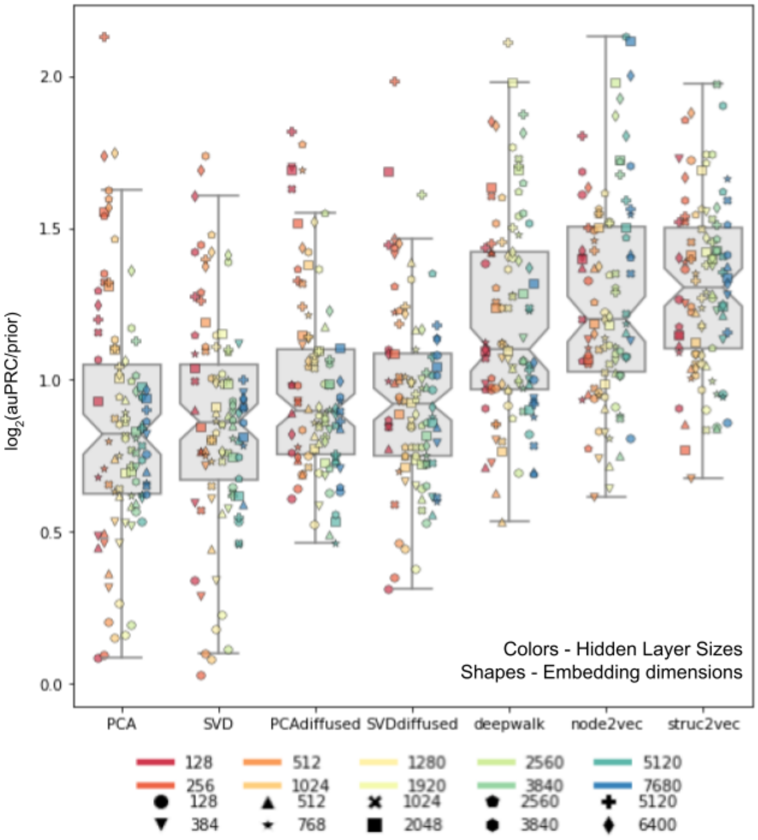

Figure 2. Performance of seven embedding techniques with varying embedding dimensions and hidden layer sizes. log2(auPRC/prior) denotes the log of the ratio of the area under the precision-recall curve (auPRC) of the real method over the auPRC expected by random chance (prior).

Fig. 2 shows the performance of the seven embedding types. The performance for each model for a hidden layer size and embedding dimension combination is the median across all tissues. The performance metric is log$_2$(auPRC/prior) indicates the fold increase of the auPRC over what is expected for a random model. This metric is well-suited for highly imbalanced tasks and is high for models where the top predicted samples are indeed true positives.

Our results clearly show that walk-based embeddings (*deepwalk*, *node2vec*, *struct2vec*) outperform traditional methods (*PCA*, *SVD*, *PCAdiffused*, *SVDdiffused*). This observation is particularly important as the original *OnClass* method used *SVDdiffusion* as the ontology embedding technique. This work demonstrates that sophisticated graph-based embedding methods create more accurate distributed representations of complex biological ontologies for the task of classifying gene expression profiles. Future extensions of our work will include conducting a similar analysis for both cell type and disease classification, as well as exploring additional embedding techniques such as elliptical embedding techniques specifically designed to faithfully capture hierarchical structures.

[1] Wang et.al., (2021) Leveraging the Cell Ontology to classify unseen cell types. Nat. Comm. **12** 5556