A Multi-species Functional Embedding Integrating Sequence and Network Structure

Mark D.M. Leiserson[1], Jason Fan[1], Anthony Cannistra[2], Inbar Fried[3], Tim Lim[4], Thomas Schaffner[5], Mark Crovella[4], and Benjamin Hescott[6]

[1] Department of Computer Science, University of Maryland, College Park
[2] Department of Biology, University of Washington
[3] University of North Carolina Medical School
[4] Department of Computer Science, Boston University
[5] Department of Computer Science, Princeton University
[6] College of Computer and Information Science, Northeastern University

Introduction. Transferring biological knowledge between species is fundamental for many important problems in genetics. These problems range from the molecular-level, such as predicting protein function or genetic interactions [4], to the organism-level, such as predicting human disease models [5]. The most common approach researchers have taken is to use orthologs inferred from DNA sequencing data. More recently, researchers have sought to expand beyond sequence-based orthologs using high-throughput proteomics data under the hypothesis that genes with similar topology in protein-protein interaction (PPI) networks have similar functions. Many methods have been introduced to infer homology across species (i.e. a node matching) from sequence similarity and PPI networks, including network alignment [1]. More recently, Jacunski, et al. [4] identified *connectivity homologous* gene pairs using a small set of features derived from PPI networks. These prior works are focused on node matching and constructing node feature vectors, but do not address the problem of embedding genes from different species into a shared, general-purpose space.

Methods. We introduce a new algorithm, H̲omology A̲ssessment across N̲etworks using D̲iffusion and L̲andmarks (HANDL), that leverages graph kernels to embed nodes from two PPI networks into a biologically meaningful and general-purpose vector space using network and sequence data.[1] Kernels, particularly kernels that capture random walks and/or heat diffusion processes on graphs, have been widely and successfully used for computing similarity between nodes within biological networks [2].

The main computational challenge HANDL solves is relating network kernel matrices from different species. Because the kernel matrices from networks of different species have different dimensions, traditional kernel transfer learning approaches (e.g. [3]) cannot be directly applied. We show a schematic of the HANDL algorithm in Figure 1. HANDL takes as input a source network, a target network, and a set of landmarks shared between the networks to embed nodes

[1] An implementation of HANDL is available at `https://github.com/lrgr/HANDL`.

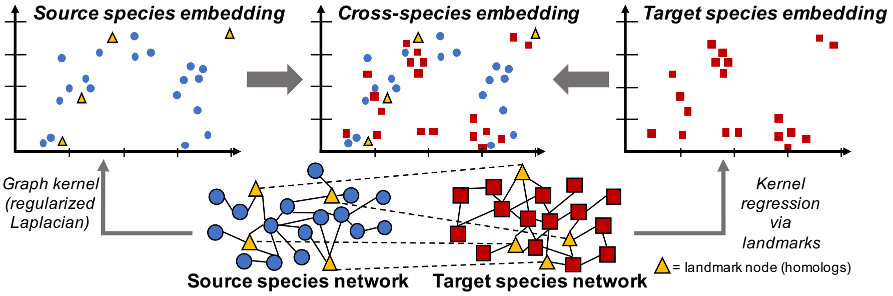

Fig. 1. HANDL embeds proteins from different species into a shared vector space

from the *target* species into the vector space of the *source* species. The inner-product between embeddings gives *HANDL similarity scores* between nodes in different species. As HANDL is a general algorithm, the landmarks and graph kernel can be customized for particular applications. In this work, we use a subset of the homologs as landmarks and the regularized Laplacian kernel specifically to capture protein functional similarity.

Results We show that the human-mouse and baker's-fission yeast cross-species embeddings constructed by HANDL are biologically meaningful with three cross-species tasks. First, we find that HANDL similarity scores are strongly correlated with cross-species functional similarity, and that pairs with the highest HANDL similarity scores are more functionally similar than pairs with the closest connectivity homology profiles [4]. Next, we use the algorithm and data from McGary, et al. [5] and *HANDL-homologs* (node pairs with high HANDL similarity scores) to find new, novel human-mouse disease models (phenologs, i.e orthologous phenotypes) that are supported by biological literature. Finally, we show that node vectors themselves are of more general use. We use HANDL to transfer knowledge of synthetic lethal (SL) interactions in baker's to fission yeast (and vice versa). We compute HANDL-embeddings for the source and target species then train a support vector machine (SVM) only on embeddings of the source species. We find that that the SVM also separates embeddings of the target species with respect to SLs and non-SLs on previously unseen data.

These results show how HANDL can transfer knowledge of genetics between humans and model organisms. We anticipate that HANDL can serve as the foundation for more sophisticated approaches for transfer learning across species.

References

1. Clark, C., and Kalita, J. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics 30*, 16 (2014), 2351–2359.
2. Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* (2017).
3. Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., and Scholkopf, B. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems* (2006), 601–608.
4. Jacunski, A., Dixon, S. J., and Tatonetti, N. P. Connectivity homology enables inter-species network models of synthetic lethality. *PLoS. Comp. Bio. 11*, 10 (2015), e1004506.
5. McGary, K. L., Park, T., Woods, J. O., Cha, H., Wallingford, J. B., and Marcotte, E. M. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. 107*, 14 (10 2010), 6544–6549.