COMBINING GRAPH AND SEQUENCE INFORMATION TO LEARN PROTEIN REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational methods that infer the function of proteins are key to understanding life at the molecular level. In recent years, representation learning has emerged as a powerful paradigm to discover new patterns among entities as varied as images, words, speech, molecules. In typical representation learning, there is only one source of data or one level of abstraction at which the learned representation occurs. However, proteins can be described by their primary, secondary, tertiary, and quaternary structure or even as nodes in protein-protein interaction networks. Given that protein function is an emergent property of all these levels of interactions in this work, we learn joint representations from both amino acid sequence and multilayer networks representing tissue-specific protein-protein interactions. Using these representations, we train machine learning models that outperform existing methods on the task of tissue-specific protein function prediction on 10 out of 13 tissues. Furthermore, we outperform existing methods by 19% on average.

1 INTRODUCTION

With few exceptions, all cells in the human body have the same genetic information. Yet the human body has more than 200 types of cells which are combined into four broad tissue categories. Proteins differences play a fundamental role in the variation among cell type, as proteins carry out various activities such as catalyzing biochemical reactions, acting as messengers, and providing structure to tissues(Creighton, 1993). Understanding how the structure of proteins relates to their function can shed light on the inner workings of cells.

Proteins are generally understood through four levels of structures (Creighton, 1993). The primary structure of a protein refers to its linear chain of amino acids. The secondary structure refers to local folded structures that form within a polypeptide due to interactions between atoms of the backbone. The most common types of secondary structures are alpha helices and beta sheets . The tertiary structure is the three dimensional shape of the protein. The protein quaternary structure is the number and arrangement of multiple folded protein subunits in a multi-subunit complex.

Understanding the relationship between these different levels of structure and the role that a protein plays is one of the grand challenges of biology. Some proteins with similar sequences play similar roles; others with high levels of sequence similarity can play different roles. To add further nuance, the same protein can play different roles depending on the tissue it is in and the state of that tissue. Recent availability of high-throughput experimental data and machine-learning based computational methods can be useful for unveiling and understanding such patterns.

In this work, we approach this problem from the angle of representation learning. Representation learning enables us to frame the problem of understanding the relationship between protein structure and function as one of embedding proteins in a vector space capturing several key properties and using such embeddings in the context of predictive tasks. From this angle, learning representations in the form of vector embeddings for proteins which are predictive of their tissue-specific function allows to train simple machine learning models mapping a given protein to its tissue-specific function.

In this work we constructed new protein representations combining different levels of abstraction. More specifically, we constructed a 128-dimensional vector for each protein where the first 64 dimensions are derived from the amino acid sequence and the remaining 64 dimensions are obtained from embedding the protein into a tissue-specific protein-protein interaction networks. Such representations are then used to train a simple linear classifier to predict tissue-specific protein function. This approach outperforms existing network-based approaches which usually only use information from the protein-protein interaction network.

The main contribution of this paper include:

- Approaching the problem of protein function prediction from the angle of representation learning using information ranging from amino acid sequence to multilayer networks including tissue-specific protein-protein interaction
- Experimentally show that such representations outperform existing methods on 10 out of 13 tissues for which we perform the experiments. The best method outperforms current ones by 19% on average. An ablation analysis that demonstrated that our state-of-the-art results are a result of the joint embeddings

2 RELATED WORK

Computational methods to predict the function of proteins fall into several categories. An important step of the pipeline is developing representations for proteins. Most existing methods focus on one level of biological abstraction and develop a representation specific to this level. For example, when looking at the primary structure, the first attempt to computationally predict the role of a protein is through sequence homology. That is, using a database of protein whose sequence and function is known, methods using string similarity will find the closest proteins and use heuristics to make a prediction based on such similarity. These methods use dynamic programming and hierarchical clustering to align multiple sequence to perform homology and find the distance of a given protein to multiple proteins stored in a database. (Feng & Doolittle, 1987) (Corpet, 1988) (Corpet, 1988) (Edgar, 2004)

Beyond sequence homology, local polypeptide chains are grouped under patterns called protein domains (Bateman et al., 2004). Protein domains evolve independently of the rest of the protein chain. They are often thought of as evolutionary advantageous building blocks which are conserved across species and proteins. The presence of such building blocks in protein is used as a proxy to infer function and protein family. Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models and has 17,929 families used to characterize unknown on the basis of motif presence.

Recently, inspired by the methods used in natural language processing, researchers have developed character-level language models by training algorithms such as long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) networks to predict the next amino acid given the previous amino acids. Many recent works have gone into training and investigating the properties learned by such language models and found that they encode many biochemical properties and can be used to recover protein families. More specifically UniRep (Alley et al., 2019) uses a multiplicative LSTM (Krause et al., 2016) trained to perform next amino acid prediction on 24 million UniRef50 (Suzek et al., 2007) amino acid sequences. The trained model is used to generate a single fixed-length vector representation of the input sequence by globally averaging intermediate mLSTM numerical summaries. SeqVec (Heinzinger et al., 2019) works by training bi-directional language model ELMo (Peters et al., 2018) on UniRef50. While such models are useful descriptors and encoders of biochemical properties, they lack the local context needed to infer protein function.

While all previously-cited methods develop representations of proteins with the basic molecular components, other methods treat proteins like social networks. Proteins rarely accomplish a function in isolation and need to bind with other proteins, in a specific tissue in a given state to accomplish a function. Using this insight, many methods describe proteins using such signals. That is, using a "guilt by association principle," they take the perspective that the role of a protein can be inferred from understanding which other proteins it interacts with (Letovsky & Kasif, 2003) (Vazquez et al., 2003) (Mostafavi et al., 2008). Representation learning methods formalizing such principles usually take as input a protein-protein interaction network represented as a graph and use methods such as matrix decomposition (Tang et al., 2011) and node embeddings (Grover & Leskovec, 2016) to develop a vector representation grouping neighboring nodes into a similar position. However, these methods do not take into account the rich information that can be learned by examining a protein's primary sequence. We aim to synthesize the previous approaches, and also take more contextual information about the tissues in which proteins interact. We use OhmNet (Zitnik & Leskovec, 2017) to include the tissue hierarchy and develop tissue-specific node embeddings taking into account local neighborhoods among proteins as well as local neighborhoods among tissues.

3 METHODS

The main idea we present is to integrate information at different levels of the biological hierarchy into the learned representation of each protein. We used information from two sources: the amino acid sequence and the tissue-specific protein-protein interaction network. We combined these representations by concatenating them into a 128 dimensional vector and trained a linear classifier to predict tissue-specific protein functions in a one vs all fashion. That is, each classifier is a binary classifier to predict if a given protein plays a given role in a specific tissue. We measure the area under the curve for each classifier and average it to have a tissue-specific AUROC.

3.1 AMINO ACID SEQUENCE REPRESENTATION

To represent the amino acid sequence, we leaned towards recent work using techniques from natural language processing to represent proteins. Indeed, recent works such as UniRep and SeqVec treat the amino acids as an alphabet and the amino acid sequence as a string in that discrete alphabet. They learn representations by leveraging the millions of protein sequences available to train a machine learning model to predict the next amino acid given the previously seen amino acids. Such training procedures have been found to develop powerful representation encompassing many known biochemical and structural properties.

For our purposes, because we wanted a 64-dimensional embedding, we used the UniRep model trained with 64 hidden units, generating a 64 dimension output vector.We also considered an alternative: SeqVec, whose output protein summary vectors are 1024 dimensions. To obtain the 64 dimension SeqVec, we included all proteins of interest in a matrix and projected them down to 64 dimensions using Principal Component Analysis.

3.2 TISSUE-SPECIFIC PROTEIN NETWORK EMBEDDING

For the second source of representation, we used the OhmNet representation. Ohmnet works by merging multiple sources of data. At the first level of hierarchy, a given tissue has a tissue-specific protein-protein interaction networks which is represented as a graph where each node is a protein and there is an edge between nodes if they are in physical contact in the tissue of interest. Then, the tissues are grouped into a tissue hierarchy which represents another directed acyclic graph. OhmNet encourages sharing of similar features among proteins with similar network neighborhoods and among proteins activated in similar tissues.

Given that the task of tissue-specific protein function is introduced in OhmNet and uses 128 dimensional vector to compare it with other methods, all of our vectors are also constructed to produce 128 dimensional vectors.

3.3 DUMMY VECTORS

To perform controlled experiments that ablate various sources of information, we constructed dummy vectors that we concatenated with either the amino acid sequence representation or the tissue-specific protein network embedding. These vectors are: Random64, a 64 dimensional random vector where each dimension is generated by sampling from a uniform distribution in the [-1,1] interval. Random128 is the corresponding 128 dimensional random vector. 0-pad, which simply pads the remaining dimensions with 0s.

Table 1: Average AUROC for tissue-specific protein function prediction

	Ohmnet128	Ohmnet64	Ohmnet-Unirep	Ohmnet-SeqVec	Ohmnet-Random	Random128
AUROC	0.52	0.52	0.54	**0.58**	0.50	0.49

4 Experiments

4.1 Experimental Setup

The goal of each experiment is to solve a multilabel binary classification problem. Each label is binary and represents a specific function (more precisely a cellular function from the Gene Ontology) in a specific tissue. On each tissue, we aim to match every active protein with zero, one or more tissue-specific functions. Using a multi-output linear classifier model, we then, for each tissue, use a separate linear classifier to predict every single protein functional activation. We evaluate and compare the protein representations from the original Ohmnet versus the augmented versions introduced in this paper. At evaluation time, the protein embeddings are splitted between training set (75%) and validation set (25%) in a randomly stratified fashion. The task at hand is to predict the unseen validation set after fitting the training set:

- Ohmnet: We use Ohmnet algorithm to learn a multi-scale network where tissues of interest are on the lowest layers. The learned protein representations are 128-dimensional vectors.
- Ohmnet64-Unirep64: We use Ohmnet algorithm to learn 64-dimensional representations of proteins in all tissues. We also learn 64-dimensional representations of said proteins from the 64 hidden layers of a 64-unit Unirep network and build a final 128 dimensional vector representation for every protein by merging these two together.
- Ohmnet64-Seqvec64: Using the previously learned Ohmnet 64-dimensional representations, we build a new set of 128-dimensional representations by mapping those embeddings with corresponding 64-dimensional sequence information learnt using SeqVec algorithm.
- Ohmnet64-Random64: Instead of adding any significant information to 64-dimensional Ohmnet representations, we pad all Ohmnet vectors with 64 dimensions of random noise generated from a uniform distribution in the open interval (-1, 1).
- Ohmnet64-0Padded: Instead of adding any significant information to 64-dimensional Ohmnet representations, we pad all Ohmnet vectors with all-zeros 64 dimensions.
- Random128: As as sanity check we also choose to evaluate the performance from completely random 128-dimensional representations.

Out of the 13 tissues we've tried:

- Ohmnet-SeqVec achieves best performance 6/13 times
- Ohmnet-Unirep achieves best performance 4/13 times
- Either therefore achieve best performance 10/13 times
- On average Ohmnet-SeqVec has a 19% higher AUROC than pure OhmNet
- On average Ohmnet-Unirep has a 13% higher AUROC than pure OhmNet

Looking at how Ohmnet-SeqVec and Ohmnet-Unirep perform shows that both Unirep and Seqvec add significant and new information that's not captured by tissue hierarchy alone. The average AUROC score from Random is a big higher than what could be expected from such representations thanks to the spikes (Placenta, Epidermis) which might also result from the huge functional class imbalance within those two tissues which, given the uniformity of the data, gets them more often than not on the right side of the hyperplane.

5 Conclusion

In this work, we have looked at how conceptually different representations of proteins could interact and complement each other for the task of predicting function. We have shown that by merging information from two task-independent representations of proteins, we make consistently better tissue-specific function predictions in 13 complex tissues. We have obtained higher scores than Ohmnet from either Ohmnet-Unirep or Ohmnet-Seqvec in 10 out of 13 tissues, with minimal increase in complexity of the classifier. Our ablation analysis demonstrates the improved results are a consequence of integrating information from different levels of the biological hierarchy.

6 Discussion/Future work

This work explores various ways of learning representations of proteins to understand protein function in its given biological context. One key takeaway is that combining representations from different level of biological abstractions leads to improved representations as judged by their ability to predict tissue-specific protein function. Recent work on developing representation from amino acid sequence enables us to take advantage of the vast amount of unlabeled sequences and work directly with proteins whether or not they have been aligned with existing sequences or annotated using known families. In the current experimental setting, we only focused on 13 tissues which had more than 2 functions and between 90 and 1400 active proteins. Further work can be done by looking at a more comprehensive set of tissues and functions. Additionally, we trained relatively simply classifiers in a one vs. all manners; more powerful approaches using complex models should naturally be explored. . We hope that our work spurs more research in representations that integrate information from multiple levels of the biological hierarchy and provide insight into the function of proteins and cells.

References

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, pp. 589333, 2019.

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22):10881–10890, 1988.

Thomas E Creighton. *Proteins: structures and molecular properties*. Macmillan, 1993.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

Da-Fei Feng and Russell F Doolittle. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, 25(4):351–360, 1987.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nachaev, Florian Matthes, and Burkhard Rost. Modeling the language of life-deep learning protein sequences. *bioRxiv*, pp. 614313, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959*, 2016.

Stanley Letovsky and Simon Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl_1):i197–i204, 2003.

Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, 9(1):S4, 2008.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Lei Tang, Xufei Wang, and Huan Liu. Scalable learning of collective behavior. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1080–1091, 2011.

Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697, 2003.

Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.