Sequence Representations and Their Utility for Predicting Protein-protein Interactions

Dhananjay Kimothi (dhananjayk@iiitd.ac.in)
Queensland University of Technology https://orcid.org/0000-0002-2444-5229
Pravesh Biyani
Indraprastha Institute of Information Technology Delhi
James M. Hogan
Queensland University of Technology
Melissa J. Davis
Walter and Eliza Hall Institute of Medical Research Bioinformatics Division

Research article

Keywords: Sequence embedding, Machine learning, Protein-Protein interactions

Posted Date: September 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-62896/v1

Version of Record: A version of this preprint was published at IEEE/ACM Transactions on Computational Biology and Bioinformatics on January 1st, 2021. See the published version at https://doi.org/10.1109/TCBB.2021.3137325.

Abstract

Background: Protein-Protein Interactions (PPIs) are a crucial mechanism underpinning the function of the cell. Predicting the likely relationship between a pair of proteins is thus an important problem in bioinformatics, and a wide range of machine-learning based methods have been proposed for this task. Their success is heavily dependent on the construction of the feature vectors, with most using a set of physicochemical properties derived from the sequence. Few work directly with the sequence itself. Recent works on embedding sequences in a low dimensional vector space has shown the utility of this approach for tasks such as protein classification and sequence search. In this paper, we extend these ideas to the PPI prediction task, making inferences from the pair instead of the individual sequences.

Methods: We propose a generic PPI prediction framework that constitutes a representation learning module for feature construction and a binary classifier. To construct the feature vector for a protein pair, we concatenate the distributed representations (embeddings) learned for the sequences of the constituent proteins. Each protein pair is represented as a 200-dimensional feature vector. To learn the embedding of a sequence, we use two established methods - Seq2Vec and BioVec, and we also introduce a novel feature construction method and call it SuperVecNW. The embeddings generated through SuperVecNW captures network information to some extent, along with the contextual information present in the sequences. Finally, we feed these feature vectors into a Random forest classifier to predict protein pair interactions.

Results: To show the efficacy of our proposed approach, we evaluate its performance on human and yeast PPI datasets, benchmarking against the established methods. Furthermore, we test our approach on three well known networks: the one-core network (CD9), the multiple-core network (Ras-Raf-Mek-Erk-Elk-Srf pathway), and the cross-connection network (Wnt-related network) and demonstrate the improvement in predicting PPIs compared to the other methods.

Conclusions: Naive low dimensional sequence embeddings provide better results on protein-protein interaction prediction task than most of the alternative representations based on other physiochemical properties. These methods require computationally modest effort due to their lower dimensionality. Advanced representation learning methods that enrich the sequence embeddings with meta information are expected to improve the results further.

Full Text

This preprint is available for download as a PDF.

Figures

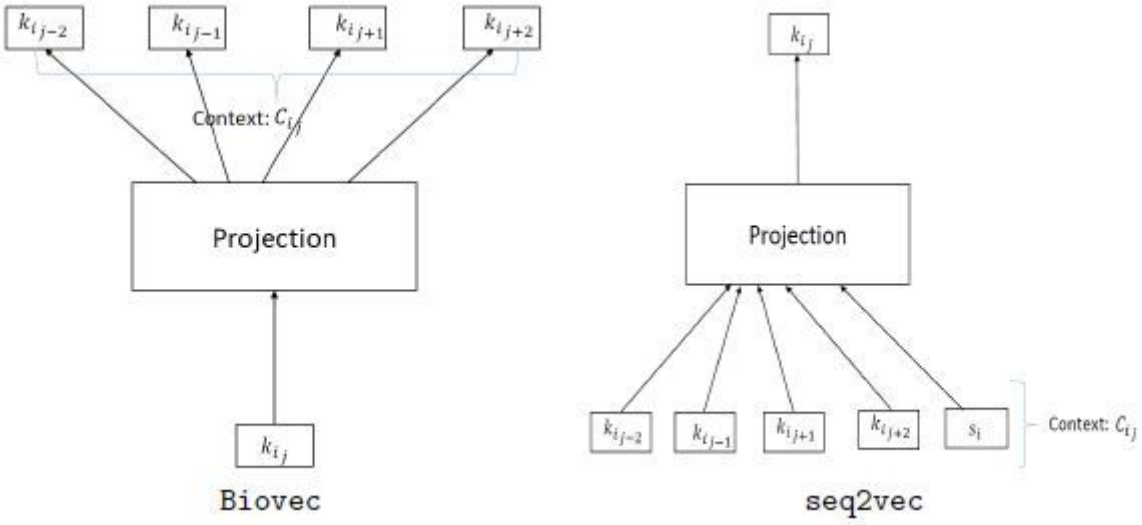

Figure 1

Architecture for BioVec and Seq2Vec models. kij represents jth k-mer in the ith sequence, si.

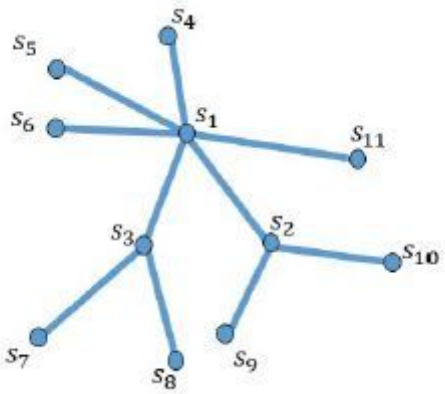

Figure 2

An example of a PPI network obtained from the training pair. Here si for i = 1, 2, ...11 represents the tag of the protein sequences. An edge between two proteins indicate that they interact, missing edge indicates the absence of known interactions.

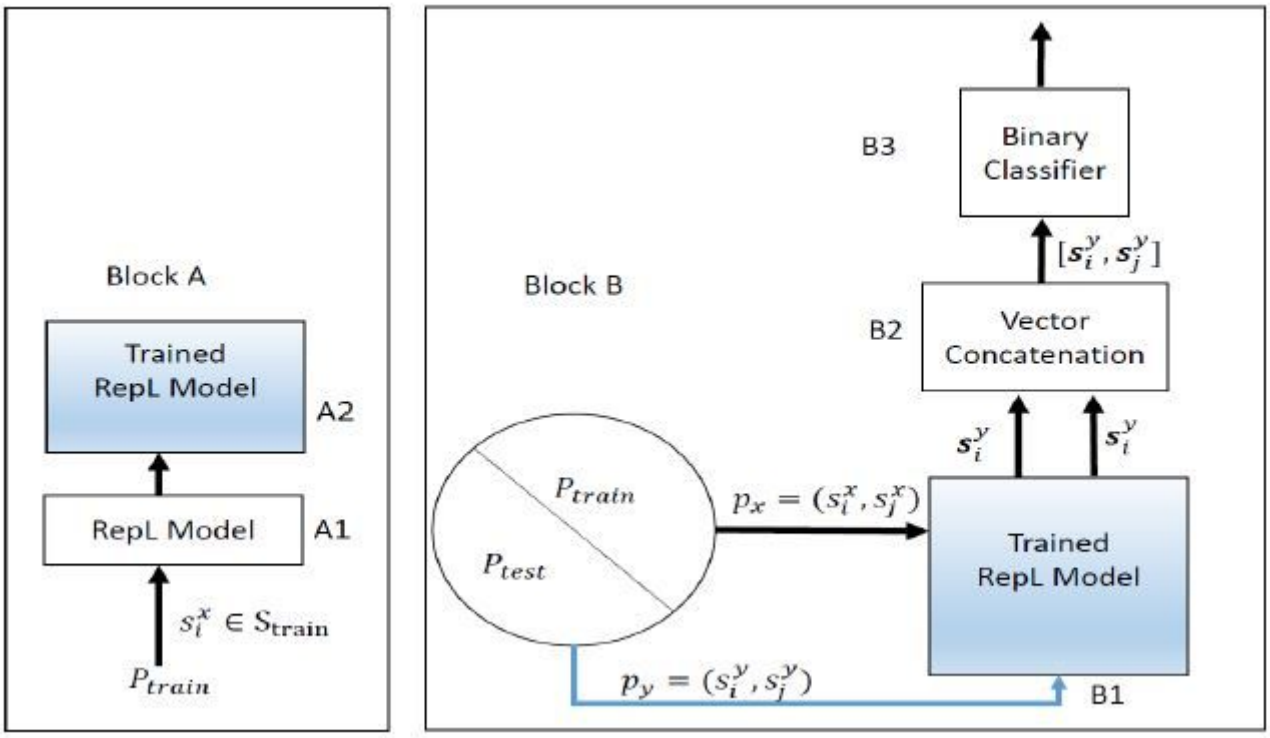

Figure 3

Proposed framework: Block A – training stage of representation learning model; Block B – constitutes a) Train/Test split of interacting/non interacting pair of protein sequences, b) vector concatenation module – for a pair of protein from Train/Test set, the vectors obtained from RL block are concatenated before finally passing to the c) binary classifier

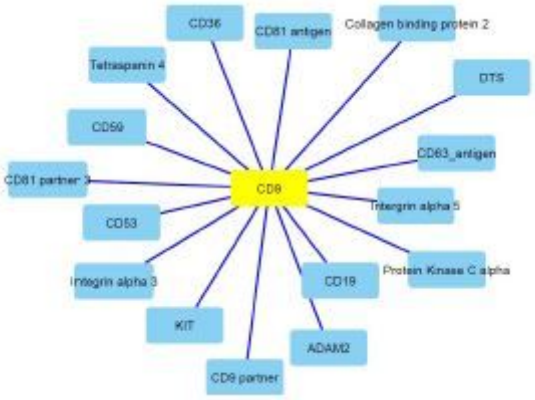

Figure 4

A one-core network for CD9

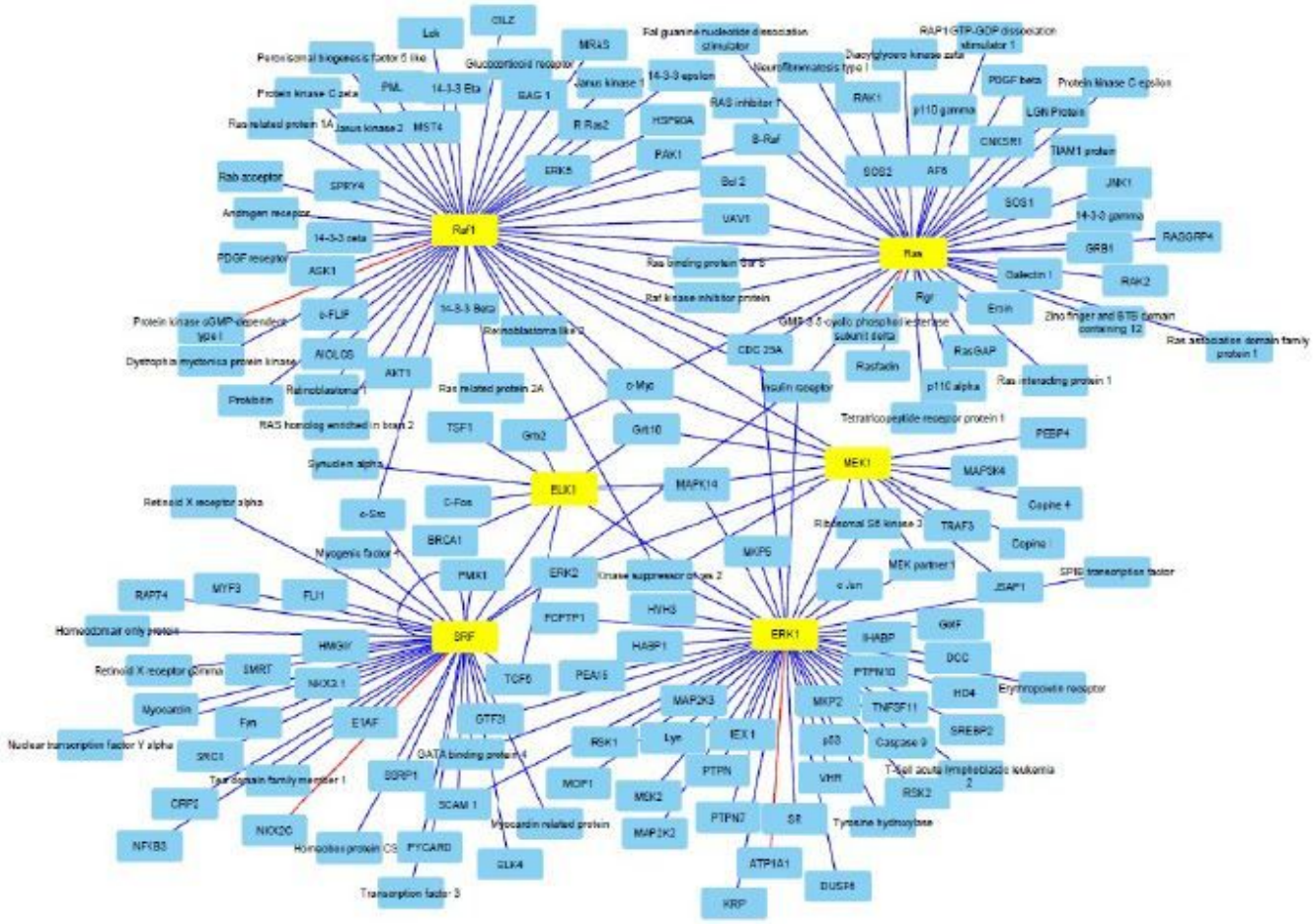

Figure 5

A multiple-core network for Ras-Raf1-Mek1-Erk1-Elk1-SRF

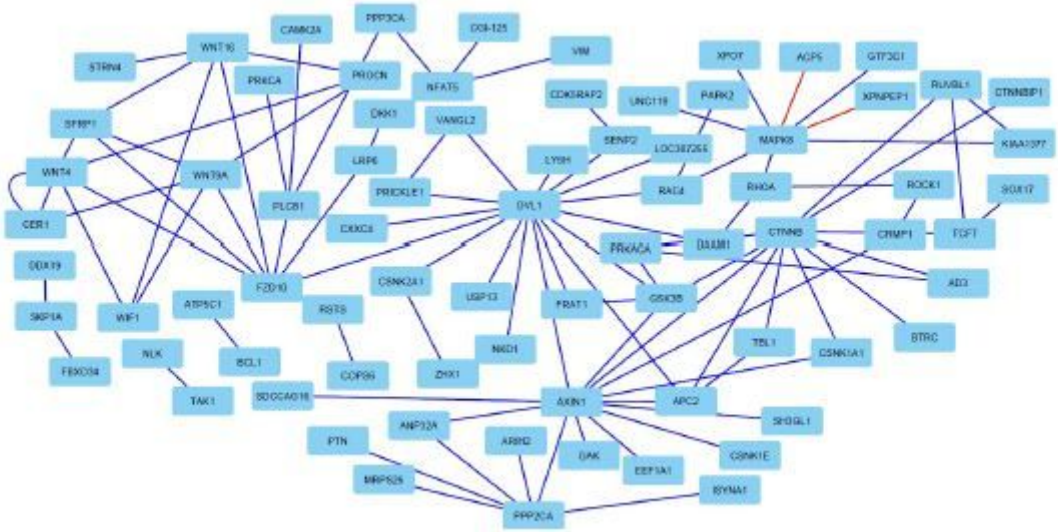

Figure 6

A crossover network for Wnt related pathway