Interpretable Structured Learning with Sparse Gated Sequence Encoder for Protein-Protein Interaction Prediction

Kishan KC[1] Feng Cui[2] Anne R. Haake[1] Rui Li[1]

[1]Golisano College of Computing and Information Sciences [2]Thomas H. Gosnell School of Life Sciences

kk3671@rit.edu kishan_kc07

Motivation

- Proteins rarely act alone as their functions tend to be regulated.
- Numerous proteins organized by their interactions forms molecular machines that carries out biological and molecular processes.
- Study of these interactions helps to:
  - Understand biological phenomenon
  - Provides insights about molecular etiology of diseases
  - Discover putative drug targets

Overview

- Assumption: Amino acid sequences contains enough information for PPI prediction.
- Protein-Protein Interaction (PPIs) prediction using variable-length amino acid sequences.
- Deep learning approaches (DPPI[1], PIPR[2]) proposed but face following challenges:
  - Black-box models and less biologically interpretable
  - Computationally expensive

- Sequence encoder with sparse and structured regularization
  - Supports interpretability with sparse gates
  - Low computation cost
  - Good prediction performance

Datasets

- Protein sequences from EMBL-EBI Reference Proteome
- Interactions from BioGRID database

Dataset	No. of proteins	No. of positive pairs	No. of negative pairs
Yeast	3651	50344	50376
Human	7028	73624	73628

Table 1: Statistics of interactions dataset from BioGRID database

Model description

- We propose interpretable deep framework, to model PPIs using variable length sequences that
  - Provides *interpretable sparsity masks.*
  - is *computationally efficient and scalable.*
  - Makes *accurate PPI predictions.*

Step 1: Sequence Encoder with Bi-GRU

- Handles variable-length sequences.
- Given a sequence $\mathbf{s}$ of length L with amino acids $[a_1, a_2, \dots, a_L]$, embedding layer projects a_l to vector representation x_l:

$$x_l = \mathbf{W}_e a_l$$

- Bidirectional GRU to learn sequential & contextualized representation of amino acids in the sequences.

$$h_l = \text{BiGRU}(x_l) = [\overrightarrow{\text{GRU}(x_l)}, \overleftarrow{\text{GRU}(x_l)}]$$

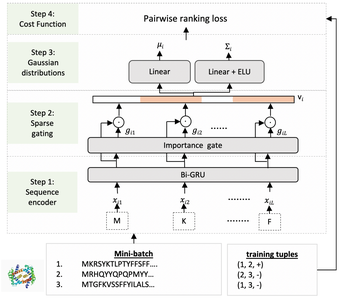

Fig 1: Block diagram of the model

Step 2: Sparsity Gating

- Not all amino acids are informative for interactions.
- Learn sparse mask to focus only on subsets of important amino acids.
- Convert h_l to score p_l:

$$p_l = \mathbf{W}_2\left(\tanh(\mathbf{W}_1 h_l + \mathbf{b}_1)\right) + \mathbf{b}_2$$

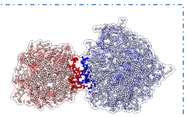

Fig 2: Protein interaction

softmax($\mathbf{p}$)	sparsemax($\mathbf{p}$)	fusedmax($\mathbf{p}$)
Full support	Sparse but distributed	Sparse and contiguous
$\frac{\exp(p_i)}{\sum_j \exp(p_j)}$	$\text{argmin}_{\{\mathbf{g} \in \Delta^K\}} \lvert\mathbf{g} - \mathbf{p}\rvert_2^2$	$\text{argmin}_{\{\mathbf{g} \in \Delta^K\}} \frac{1}{2}\left\lvert\mathbf{g} - \frac{\mathbf{p}}{\gamma}\right\rvert_2^2 + \lambda \sum_{j=1}^{L-1} \lvert\mathbf{g}_{j+1} - \mathbf{g}_j\rvert$

Step 3: Gaussian representation

- Proteins interacts with various proteins having diverse functions and different sequence patterns.
- Such diverse information can be reflected in the uncertainty of the representation.
- Protein sequence $\mathbf{s}$ is encoded to d-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

Step 4: Pairwise ranking loss

- Minimize the statistical distance between interacting proteins while maximizing the distance for non-interacting proteins
- Wasserstein distance between Gaussian representation of sequences:

$$E_{ij} = \text{Wasserstein distance}\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)\right) = \lvert\mu_i - \mu_j\rvert_2^2 + \left\lvert\Sigma_i^{\frac{1}{2}} - \Sigma_j^{\frac{1}{2}}\right\rvert_F^2$$

- Employ square-exponential loss to learn from known interactions

$$\mathcal{L} = \sum_i \sum_{(i,j) \in \mathbf{Y}^+} \sum_{(i,k) \in \mathbf{Y}^-} \left(E_{ij}^2 + \exp(-E_{ik})\right)$$

Results

- Our proposed model shows superior performance than state-of-the-art methods.

Method		Data	Yeast AUROC	Yeast AP	Human AUROC	Human AP
DPPI [7]		Profiles	0.891±0.004	0.857±0.007	0.870±0.004	0.835±0.005
PIPR [8]		sequences	0.909±0.003	0.912±0.004	0.878±0.002	0.882±0.003
Our method (sparsemax)	Ranking	Profiles	0.882±0.003	0.888±0.002	0.884±0.003	0.893±0.004
		Sequences	0.901±0.002	0.904±0.002	0.881±0.002	0.889±0.001
	Random Forest	Profiles	0.908±0.002	0.913±0.003	**0.891 ±0.005***	**0.896±0.005***
		Sequences	**0.924±0.002***	**0.925±0.001***	0.887±0.002	0.894±0.001
Our method (fusedmax)	Ranking	Profiles	0.882±0.006	0.885±0.006	0.873±0.09	0.881±0.01
		Sequences	0.898±0.001	0.900±0.002	0.874±0.002	0.883±0.001
	Random Forest	Profiles	0.906±0.004	0.912±0.005	0.872±0.015	0.877±0.015
		Sequences	0.919±0.003	0.921±0.002	0.881±0.002	0.886±0.001

Table 2: Comparison with the state-of-the-art models

- Does sparsity gating mechanism improve the performance on interaction prediction?

Model configuration		AUROC	AP
No gating		0.880±0.001	0.875±0.003
Point + RF	Softmax	0.881±0.001	0.877±0.001
	Fusedmax	0.909±0.001	0.912±0.002
	Sparsemax	0.913±0.001	0.916±0.002
Gaussian + RF	Softmax	0.882±0.001	0.879±0.002
	Fusedmax	0.919±0.003	0.921±0.001
	Sparsemax	**0.924±0.002**	**0.925±0.001**

Table 3: Study of model components on Yeast dataset

- Does learned sparsity mask match biological knowledge?

Dataset	Selected amino acids (%)	Alignment with motifs (%)
Yeast	19.24	59.05
Human	23.33	65.63

Table 4: Alignment of sparse mask with motifs

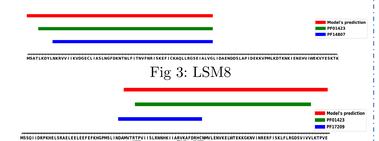

Fig 3: LSM8

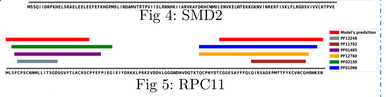

Fig 4: SMD2

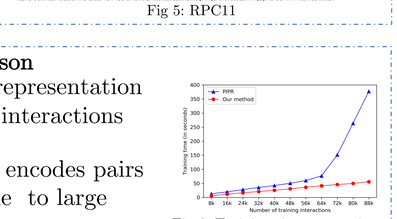

Fig 5: RPC11

- Average training time comparison
  - Encode all sequences to their representation and optimize based on known interactions
  - Other methods (DPPI, PIPR) encodes pairs of sequences and is not scalable to large number of interactions.

Fig 6: Training time comparison

References

1. Hashemifar, Somaye, et al. "Predicting protein–protein interactions through sequence-based deep learning." *Bioinformatics* 34.17 (2018): i802-i810.
2. Chen, Muhao, et al. "Multifaceted protein–protein interaction prediction based on Siamese residual rcnn." *Bioinformatics*35.14 (2019): i305-i314.

Acknowledgements

This work was supported by the NSF [NSF-1062422 to A.H.], [NSF-1850492 to R.L.] and the NIH [GM116102 to F.C.]