RIT

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

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

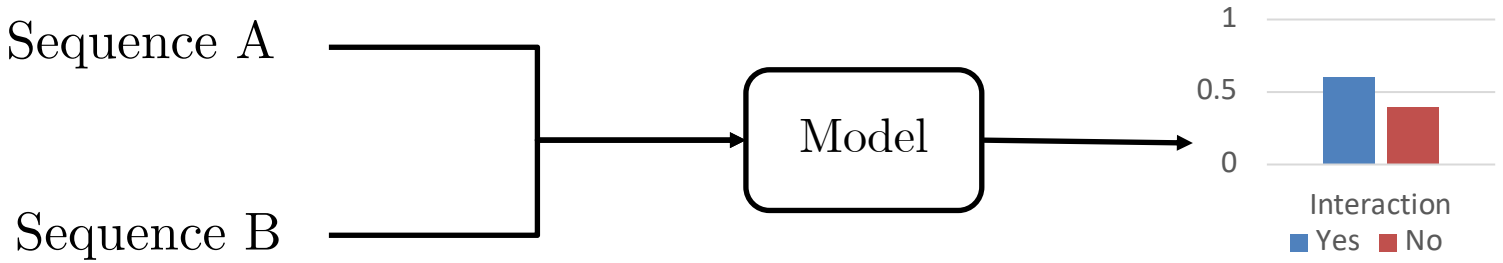
Kishan KC, Feng Cui, Anne Haake, Rui Li
Lab of Use-Inspired Computational Intelligence
Golisano College of Computing and Information Sciences
Rochester Institute of Technology (RIT), New York, USA

Background

- 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:
  - Understand biological phenomenon.
  - Insights about molecular etiology of diseases.
  - Discovery of putative drug targets.

Problem

- **Goal**: Predict the interaction between proteins from sequences



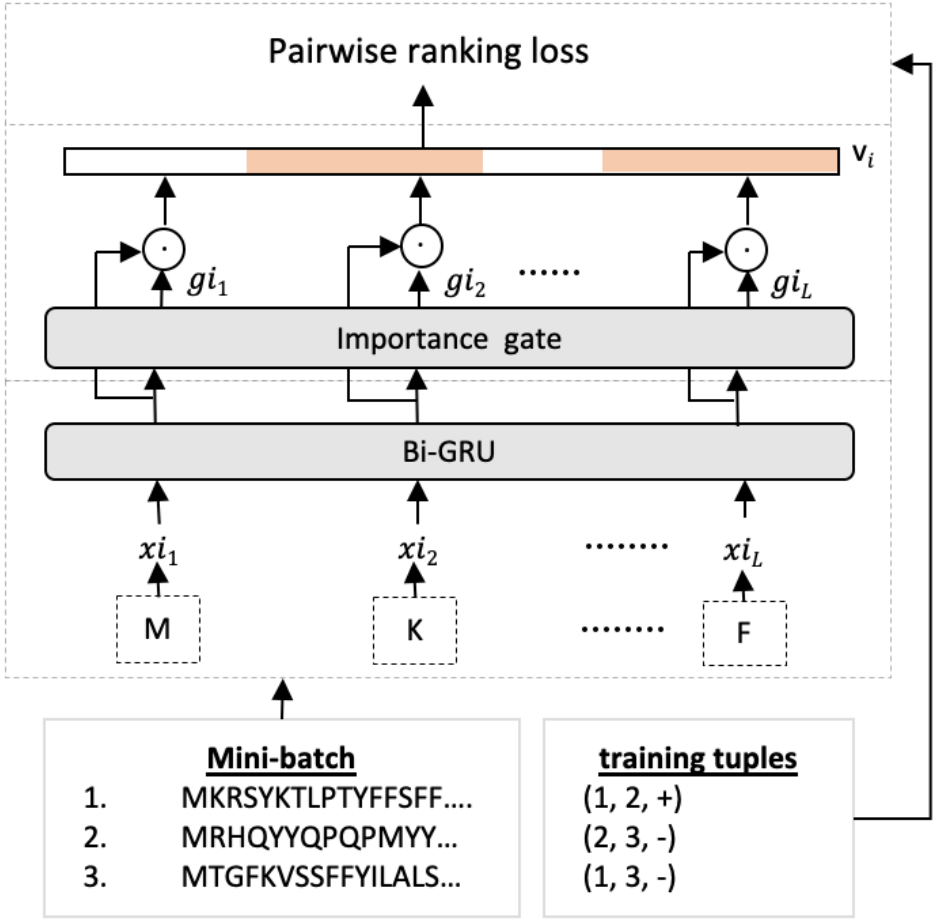
- Challenges

Lack interpretability

Computationally expensive

Proposed Method

- 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.



Sequence Encoder

- Handles variable-length sequences.
- Embedding layer projects one hot encoding a_l to vector x_l:

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

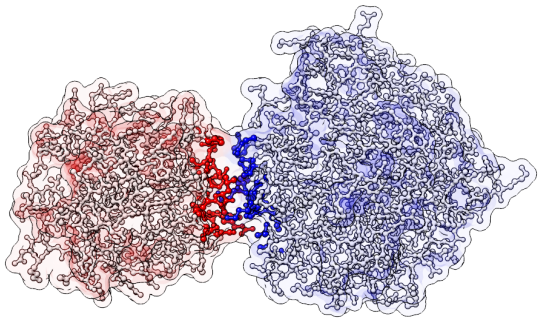
- Bidirectional GRU to learn sequential & contextualized representation of amino acids in the sequences.

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

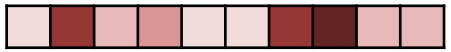
where $\overrightarrow{\text{GRU}(x_l)}$ is the forward encoding process from position 1 to L and $\overleftarrow{\text{GRU}(x_l)}$ is the backward encoding process from position L to 1.

Sparse Importance Gate

- 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:



$$p_l = \mathrm{MLP}(h_l)$$

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

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 $\mathcal{N}(\mu, \Sigma)$.

Pairwise Ranking Loss

- Minimize the statistical distance E_{ij} between interacting proteins while maximizing the distance for non-interacting proteins

$$E_{ij\,(interacting)} < E_{ik\,(non-interacting)}$$

- Wasserstein distance between Gaussian representation of sequences:

$$E_{ij} = \text{Wasserstein distance}\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}\left(\mu_j, \Sigma_j\right)\right) = \left|\mu_i - \mu_j\right|_2^2 + \left|\Sigma_i^{\frac{1}{2}} - \Sigma_j^{\frac{1}{2}}\right|_F^2$$

- 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

- Datasets

Data	No. of proteins	No. of positive pairs	No. of negative pairs
Yeast	3,651	50,344	50,376
Human	7,028	73,624	73,628

Table: Datasets used for PPI prediction

- Our proposed method performs better than state-of-the-art methods.

Method	Classifier	Yeast		Human	
		AUROC	AP	AUROC	AP
Our method (sparsemax)	Ranking	0.901±0.002	0.904±0.002	0.881±0.002	0.889±0.001
	Random Forest	**0.924±0.002***	**0.925±0.001***	**0.887±0.002***	**0.894±0.001***
Our method (fusedmax)	Ranking	0.898±0.001	0.900±0.002	0.874±0.002	0.883±0.001
	Random Forest	0.919±0.003	0.921±0.002	0.881±0.002	0.886±0.001
DPPI		0.891±0.004	0.857±0.007	0.870±0.004	0.835±0.005
PIPR		0.909±0.003	0.912±0.004	0.878±0.002	0.882±0.003

Table: Average AUROC and AP scores for PPI prediction

Ablation study

- 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: Study of sparse gates on Yeast datasets

Evaluating sparse gates

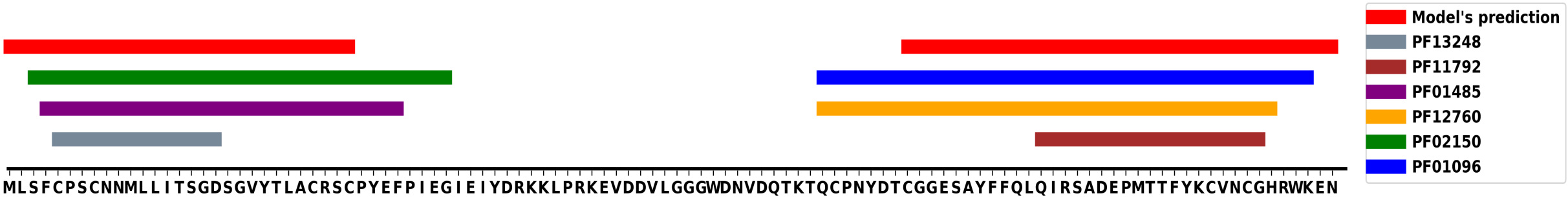
- Does learned sparsity mask match biological knowledge?

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

Interpretability

- Visualization

LSM8

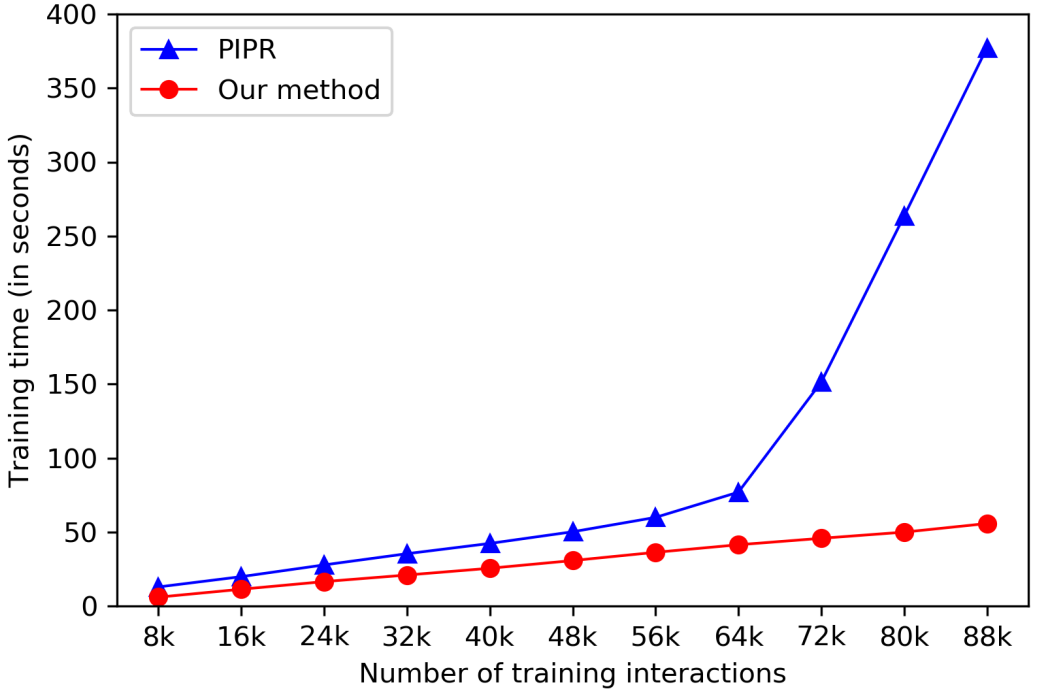


RPC11

Efficiency

- 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.



Conclusion

- We propose deep framework to model and predict PPIs using variable length sequences
  - is computationally efficient and scalable.
  - Makes accurate PPI predictions.
  - Learns sparse masks to provide interpretability.

Acknowledgements

Funding

Thanks