GRADUATE SHOWCASE
RESEARCH | CREATE | INNOVATE

Learning Sparse and Structured Gaussian Embedding of protein sequences using pairwise constraints

Kishan K C | PhD, GCCIS

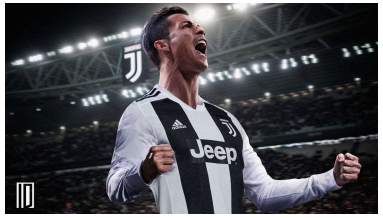
November 22, 2019

OFFICE OF GRADUATE EDUCATION

RIT

Introduction

- What do they have in common?
  - Basic biology to keep them alive and functioning.
- E.g. Undergoing several different biochemical processes such as:
  - Breaking down food
  - Repairing tissues or worn out cells
  - Replicating DNA



YOU

Proteins

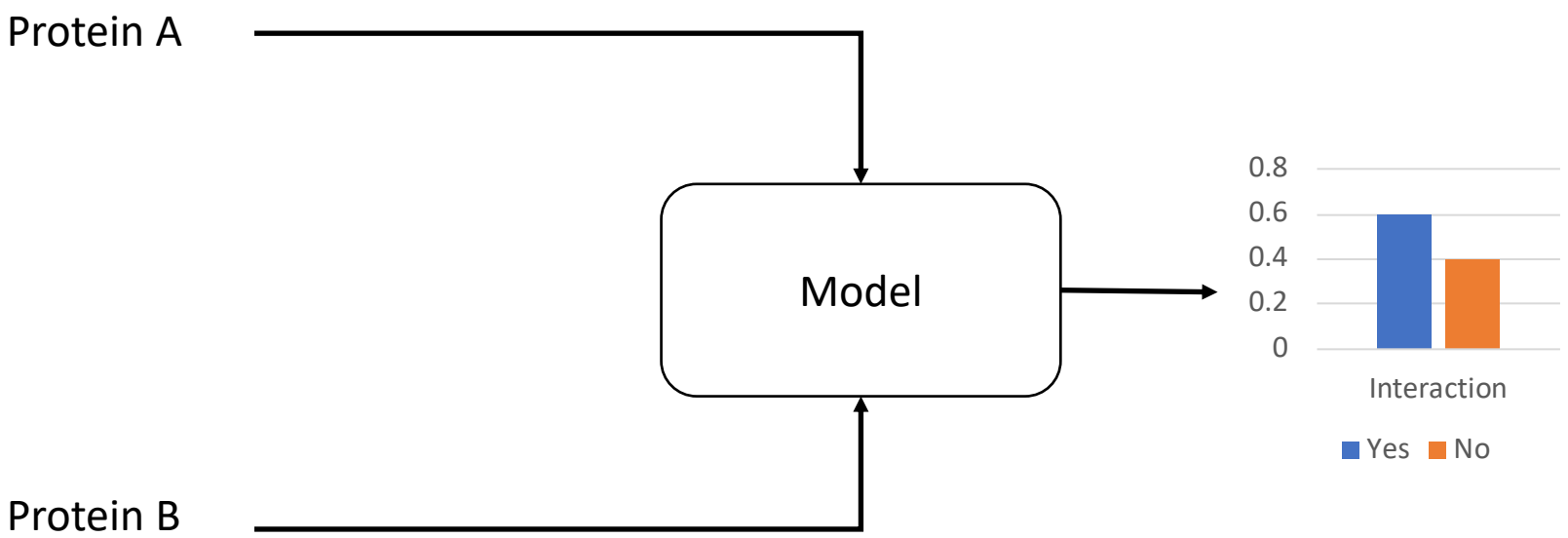
- Proteins allow organisms to undergo these basic life processes
- Need energy from your last food?
  - Proteins build the enzymes used by the digestive system to break down and extract nutrients from food.
- Want to build muscles?
  - Muscles are build from proteins.
- How can organism stay alive?
  - Proteins form the enzymes need to replicate DNA and replace old and worn out cells.

Protein-Protein interactions (PPI)

- Proteins rarely act alone as their functions tend to be regulated
- Numerous proteins organized by their physical contacts forms molecular machines that carries out biological and molecular processes
- Study of these contacts:
  - Understand biological phenomenon
  - Insights about molecular etiology of diseases
  - Discovery of putative drug targets
- Contacts between proteins: Protein Protein interactions (PPI)

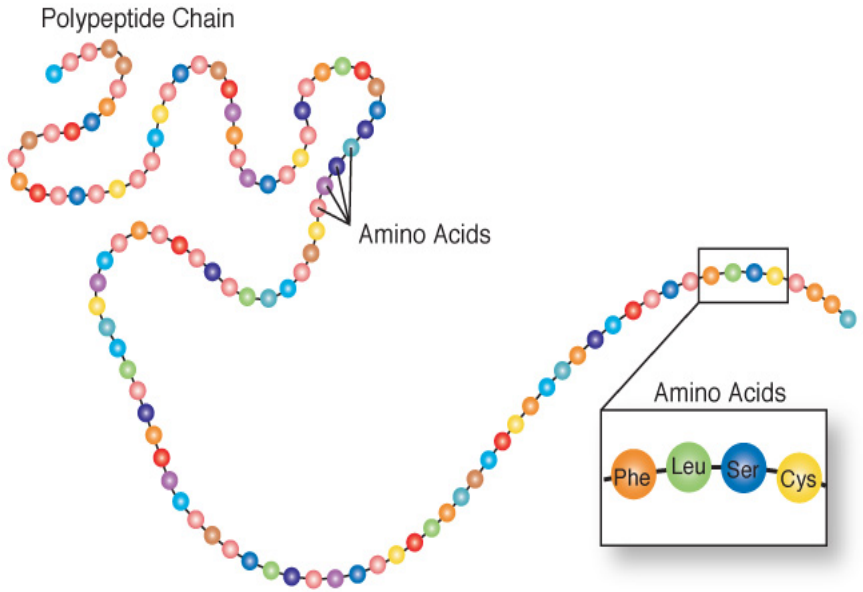
Problem

Predict if two proteins interact.



Amino acid sequence

- Proteins are made up of smaller units called amino acids.
- Strings of amino acids are arranged in particular order.
  - Protein **A5Z2X5**
    MRPAQLLLNTAKKTSGGYKIPVELTPLFLAVGVALCSGTYFT
    YKKLRTDETLRLTGNPEL SSLDEVLAKDKD
- Amino acid sequence is the primary structure of the protein.
  - determines the protein's unique three-dimensional shape.



Amino Acids

Ala: Alanine	Gln: Glutamine	Leu: Leucine	Ser: Serine
Arg: Arginine	Glu: Glutamic acid	Lys: Lysine	Thr: Threonine
Asn: Asparagine	Gly: Glycine	Met: Methionine	Trp: Tryptophane
Asp:Aspartic acid	His: Histidine	Phe: Phenylalanine	Tyr: Tyrosisne
Cys:Cysteine	Ile: Isoleucine	Pro: Proline	Val: Valine

rarediseases.info.nih.gov/GlossaryDescription/14/0

Previous works

- Predict interactions between a pair of protein sequences
- State-of-the-art methods proposed Siamese network to model the mutual influence between proteins.

DPPI (Hashemifar et al. 2018)

- Deep convolutional neural network (CNN) to learn protein representation
- Doesn't consider sequential information of amino acids

PIPR (Chen et al. 2019)

- Deep Recurrent Convolutional neural network (RCNN) to learn protein representation

Challenges

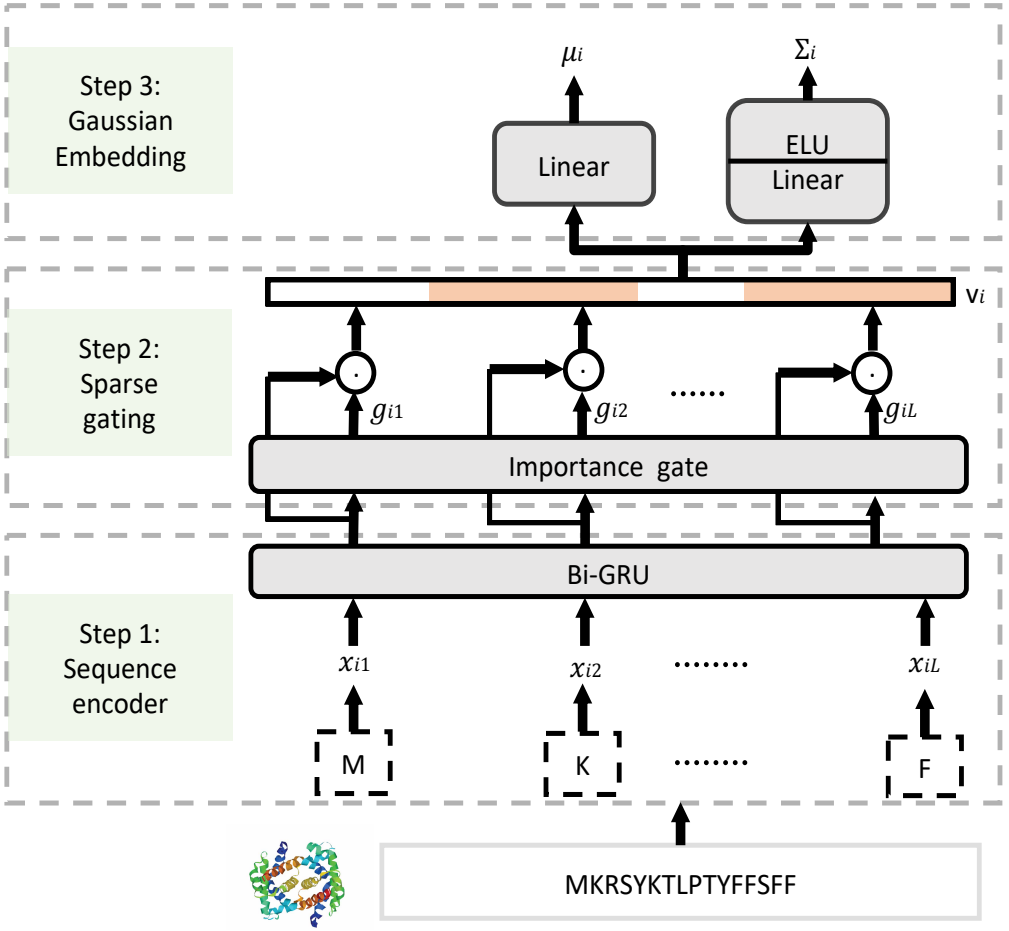
- Hard to explain the predictions i.e. lack transparency
- Computationally expensive approach in Siamese setting

  For instance:

  - Human has nearly 20,000 proteins.
  - Nearly 200 million possible interactions.
  - If processing an interaction takes 1 second, total processing time > 6 years.

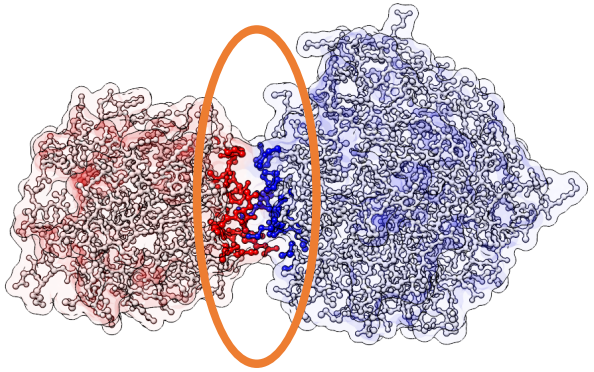
Proposed approach

- Sequence encoder
  - Bidirectional GRU to model contextual and sequential properties of amino acids
  - Handles variable length sequences
  - Captures long term dependencies
- Sparse gating
  - Guides model to selectively focus on specific amino acids in the sequence
- Gaussian embedding
  - Model the uncertainty about the representation of amino acid sequences



Sparse gating mechanism

- Proteins interact via interface, small region of protein structure



Softmax	Sparsemax[1]	Fusedmax[2]
• Full support	• Sparse weight but distributed	• Sparse and contiguous

MRPAQLLLNTAKKTSGGYKIPVELTPLFLAV
GVALCSGTYFTYKKLRTDETLRLTGNPEL
SSLDEVLAKDKD

→ Sparse Gating →

MRPAQLLLNTAKKTS**GGYKIPVELTPLFLA**
VGVALCSGTYFTYKKLRTDETLRLTGNPEL
SSLDEVLAKDKD

1. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification, André F. T. Martins, Ramón Fernandez Astudillo, ICML 2016.
2. A Regularized Framework for Sparse and Structured Neural Attention, Niculae, Vlad, and Mathieu Blondel, NeurIPS 2017.

Experimental setup

- Select a batch of n protein sequences
- Encode these sequences to Gaussian distributions
- Retrieve positive and negative interactions that involve these n proteins
- Minimize the statistical distance between interacting proteins while maximizing the distance for noninteracting proteins.

$$dist^2 = ||\mu_i - \mu_j||_2^2 + ||\Sigma_i^{\frac{1}{2}} - \Sigma_j^{\frac{1}{2}}||_F^2$$

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

Results

Method	Yeast		Human	
	AUROC	AP	AUROC	AP
Our method + sparsemax	**0.924±0.002**	**0.925±0.001**	**0.887±0.003**	**0.891±0.002**
Our method + fusedmax	0.919±0.003	0.921±0.002	0.881±0.002	0.886±0.001
DPPI (Hashemifar et al. 2018)	0.891±0.004	0.857±0.007	0.870±0.004	0.835±0.005
PIPR (Chen et al. 2019)	0.909±0.003	0.912±0.004	0.878±0.002	0.882±0.003

Table 1: Comparison with the state-of-the-art models

Ablation study

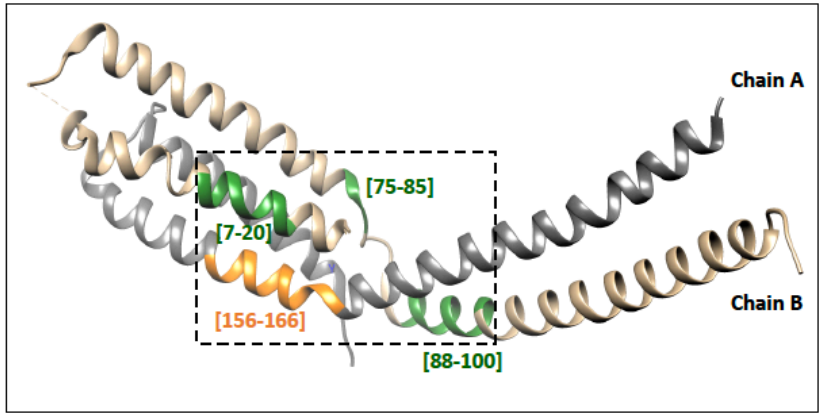
Model configuration		AUROC	AP
No gating		0.880±0.001	0.875±0.003
SE + RF	Softmax	0.881±0.001	0.877±0.001
	Fusedmax	0.909±0.001	0.912±0.002
	Sparsemax	0.913±0.001	0.916±0.002
GE + RF	Softmax	0.882±0.001	0.879±0.002
	Fusedmax	0.919±0.003	0.921±0.001
	Sparsemax	**0.924±0.002**	**0.925±0.001**

Table 2: Study of model components on Yeast dataset

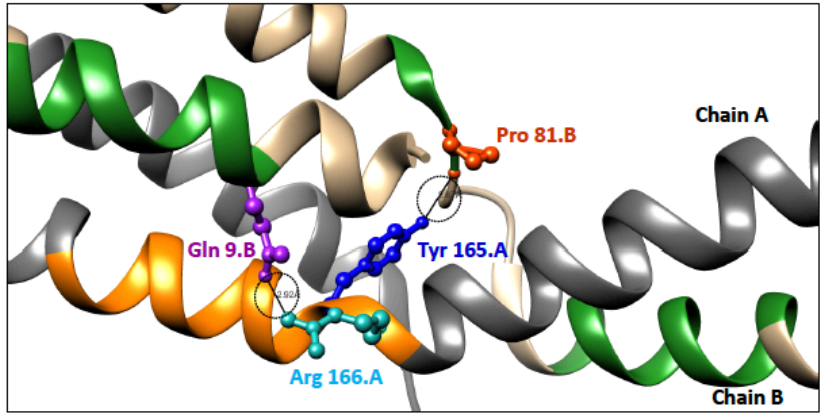
Interpretability

	Ground truth motif	Sparsemax	Fusedmax
LSM8	MSATLKDYLNKRVVIIKVDG ECLIASLNGFDKNTNLFITN VFNRISKEFICKAQLLRGSE IALVGLIDAENDDSLAPIDE KKVPMLKDTKNKIENEHVIW EKVYESKTK	MSATLKDYLNKRVVIIKVDG ECLIASLNGFDKNTNLFITN VFNRISKEFICKAQLLRGSE IALVGLIDAENDDSLAPIDE KKVPMLKDTKNKIENEHVIW EKVYESKTK	MSATLKDYLNKRVVIIKVDG ECLIASLNGFDKNTNLFITN VFNRISKEFICKAQLLRGSE IALVGLIDAENDDSLAPIDE KKVPMLKDTKNKIENEHVIW EKVYESKTK
SMD2	MSSQIIDRPKHELSRAELEE LEEFEFKHGPMSLINDAMVT RTPVIISLRNNHKIIARVKA FDRHCNMVLENVKELWTEKK GKNVINRERFISKLFLRGDS VIVVLKTPVE	MSSQIIDRPKHELSRAELEE LEEFEFKHGPMSLINDAMVT RTPVIISLRNNHKIIARVKA FDRHCNMVLENVKELWTEKK GKNVINRERFISKLFLRGDS VIVVLKTPVE	MSSQIIDRPKHELSRAELEE LEEFEFKHGPMSLINDAMVT RTPVIISLRNNHKIIARVKA FDRHCNMVLENVKELWTEKK GKNVINRERFISKLFLRGDS VIVVLKTPVE
RPC10	MPPLPQNYAQQQPSNWDKFK MGLMMGTTVGVCTGILFGGF AIATQGPGPDGVVRTLGKYI AGSAGTFGLFMSIGSIIRSD SESSPMSHPNLNLQQQARLE MWKLRAKYGIRKD	MPPLPQNYAQQQPSNWDKFK MGLMMGTTVGVCTGILFGGF AIATQGPGPDGVVRTLGKYI AGSAGTFGLFMSIGSIIRSD SESSPMSHPNLNLQQQARLE MWKLRAKYGIRKD	MPPLPQNYAQQQPSNWDKFK MGLMMGTTVGVCTGILFGGF AIATQGPGPDGVVRTLGKYI AGSAGTFGLFMSIGSIIRSD SESSPMSHPNLNLQQQARLE MWKLRAKYGIRKD
MGR2	MLSFCPSCNNMLLITSGDSG VYTLACRSCPYEFPIEGIEI YDRKKLPRKEVDDVLGGGWD NVDQTKTQCPNYDTCGGESA YFFQLQIRSADEPMTTFYKC VNCGHRWKEN	MLSFCPSCNNMLLITSGDSG VYTLACRSCPYEFPIEGIEI YDRKKLPRKEVDDVLGGGWD NVDQTKTQCPNYDTCGGESA YFFQLQIRSADEPMTTFYKC VNCGHRWKEN	MLSFCPSCNNMLLITSGDSG VYTLACRSCPYEFPIEGIEI YDRKKLPRKEVDDVLGGGWD NVDQTKTQCPNYDTCGGESA YFFQLQIRSADEPMTTFYKC VNCGHRWKEN

Interpretability: case study



(a) Important segments predicted by our model.



(b) Validated contact between the residues in the predicted segments.

RIT