Graph Alignment-Based Protein Comparison

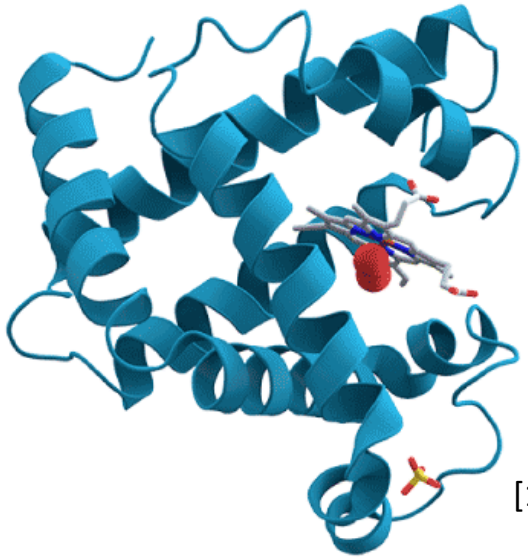
Daniel Xu, Tanisha Saxena

MIT PRIMES

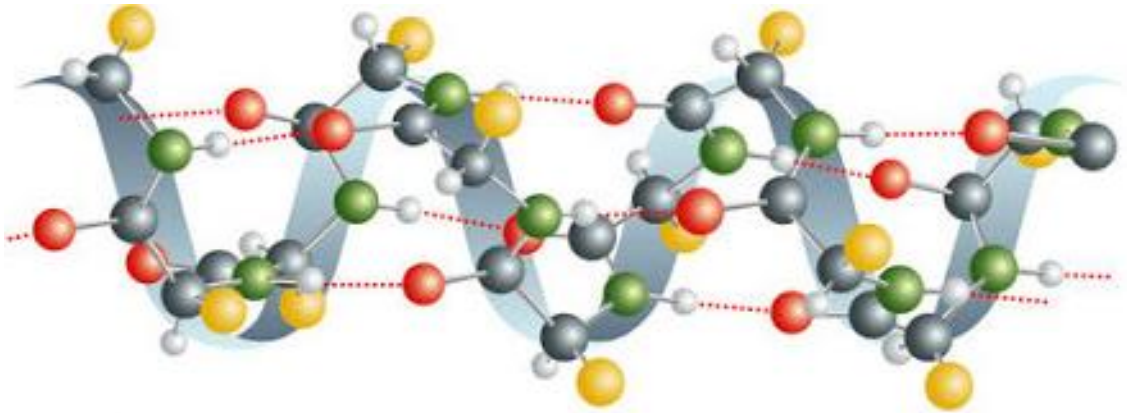
Mentor: Younhun Kim

Background

Proteins are chains of **amino acids** that control many functions within living cells



[1]



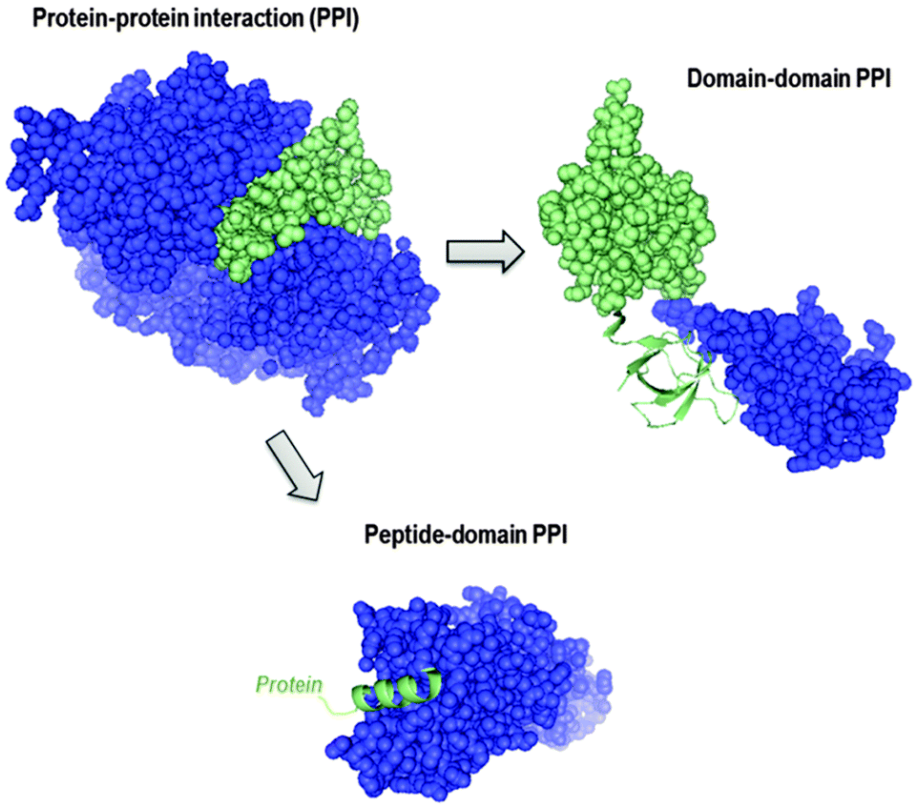
[2]

A **protein-protein interaction** is the physical contact between two protein molecules, localized at the binding sites.

[1] Source: BYJU'S The Learning App
[2] Advameg Biology Reference

Motivation

Which proteins in the human body do COVID-19 proteins interact with?



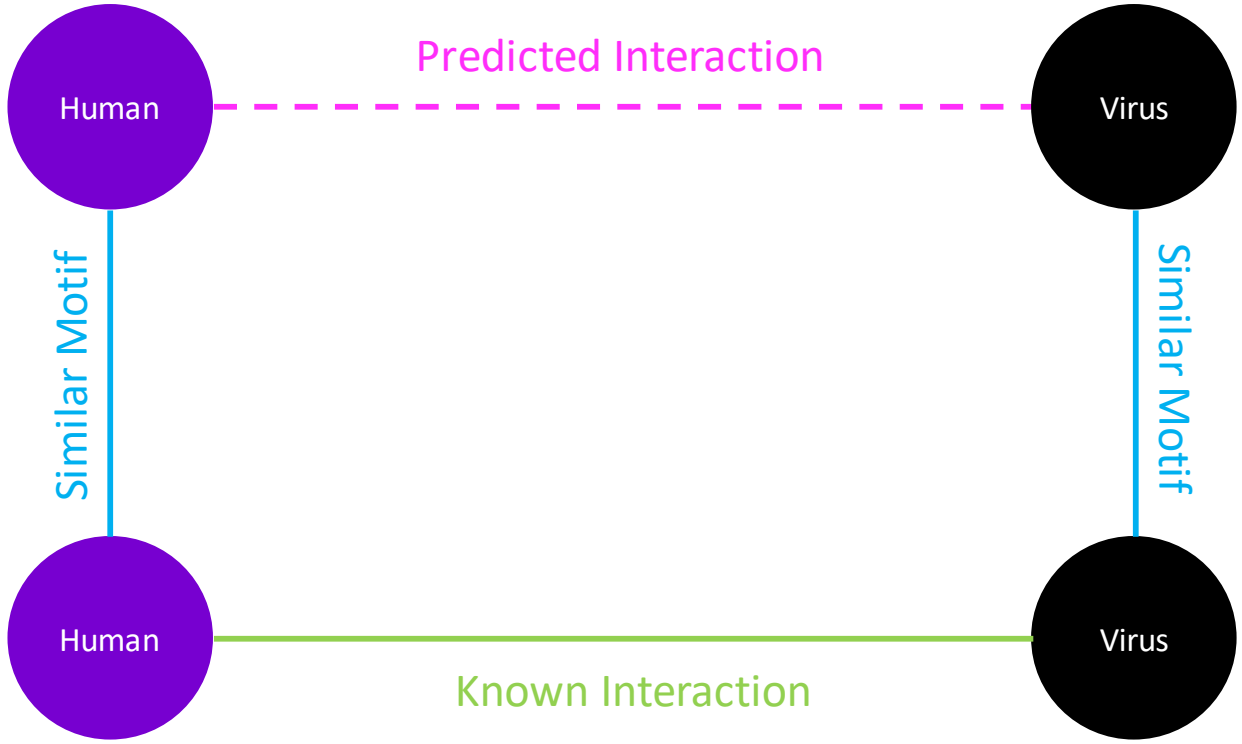
Source: Modulating Protein-Protein Interactions: The Potential of Peptides

A First Attempt

We assume that similar proteins act in similar ways

When creating our methods, we kept this assumption in mind to determine which direction made the most sense

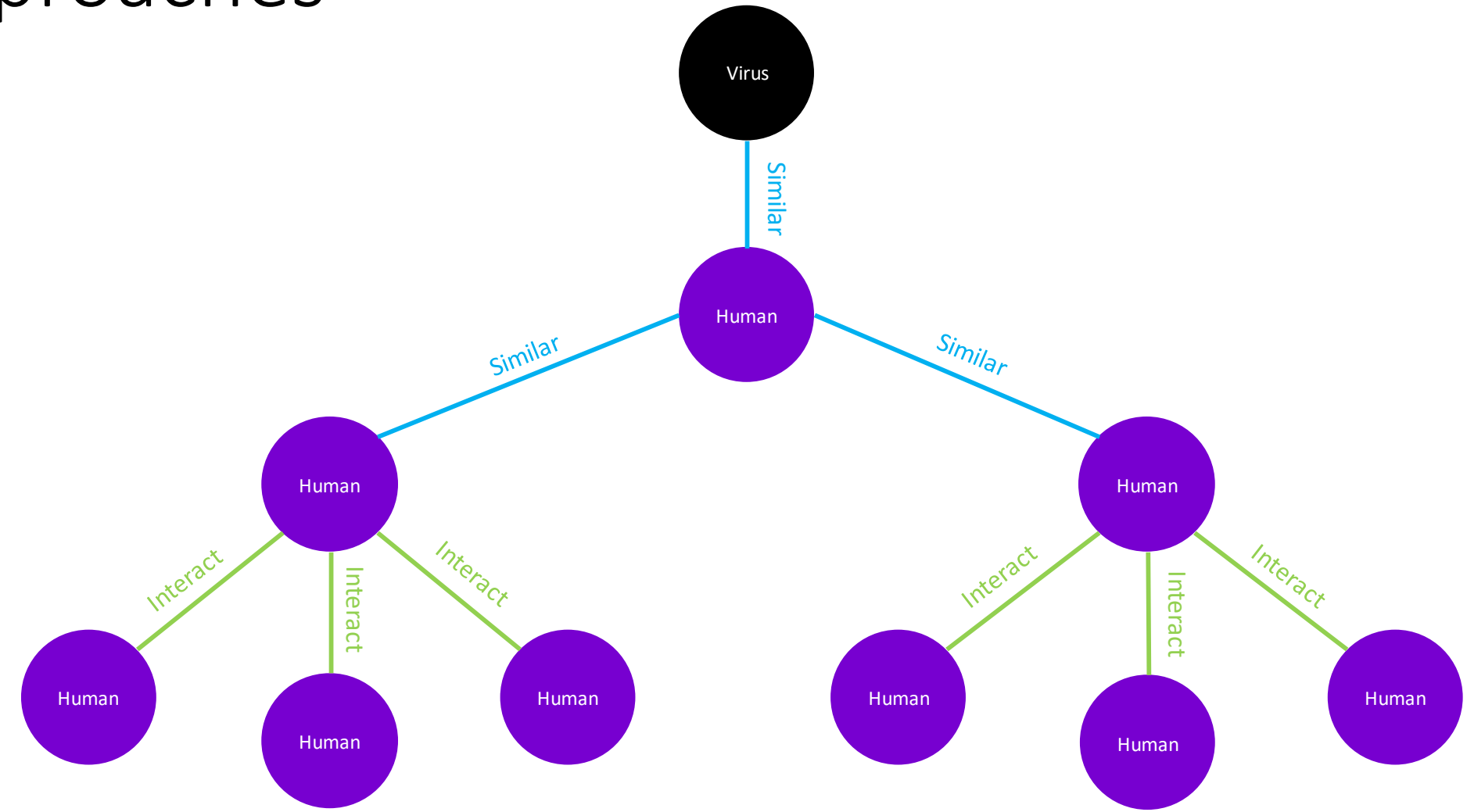
A First Attempt



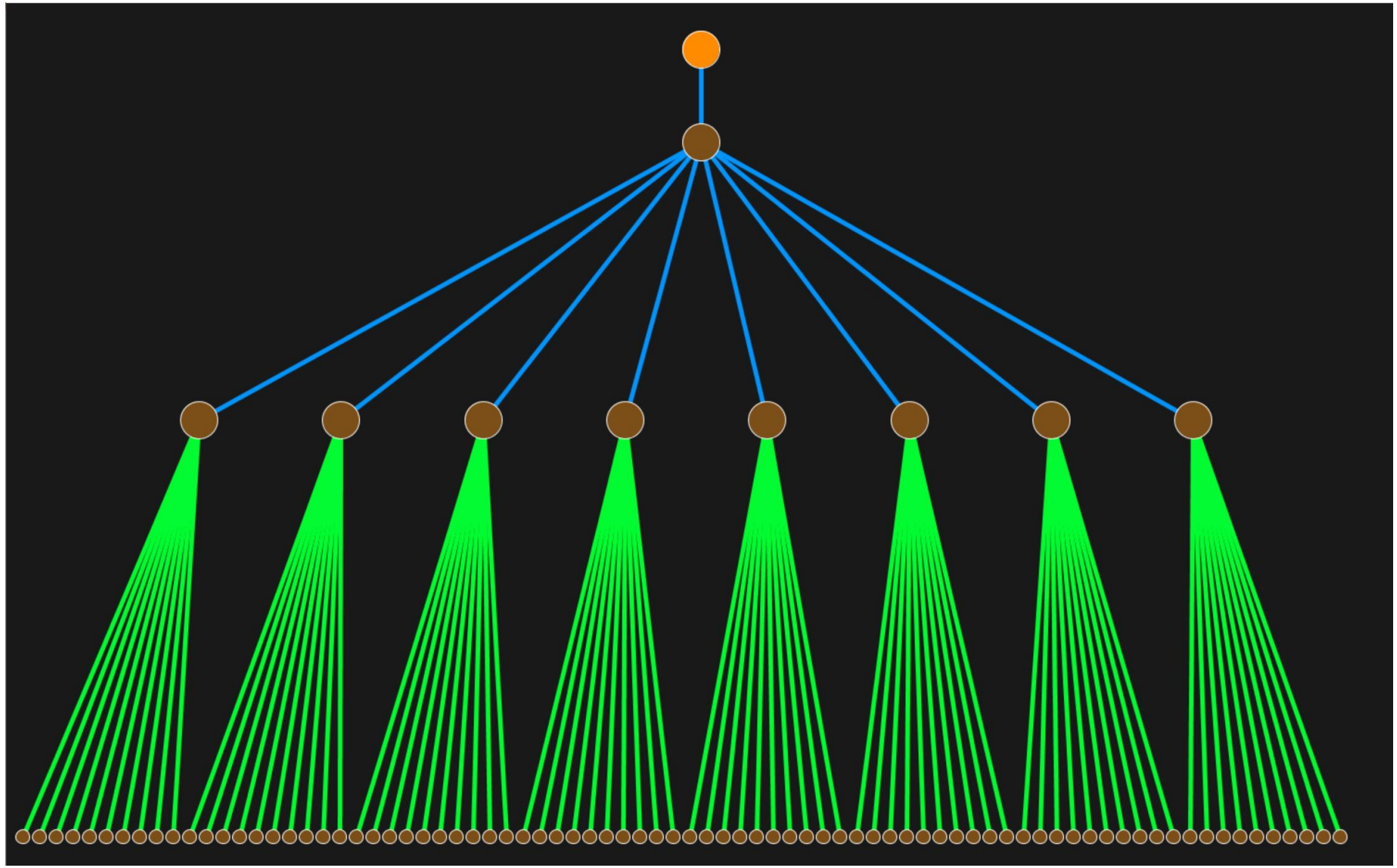
Data

- Eukaryotic Linear Motif (ELM)
  - Input: amino acid sequence
  - Output: list of proteins with similar motifs
- STRING
  - Input: protein by name or sequence
  - Output: top 10 most interactive proteins with the given one
- National Center for Biotechnology Information (NCBI)
  - Allows one to search for the sequence of a protein

Approaches

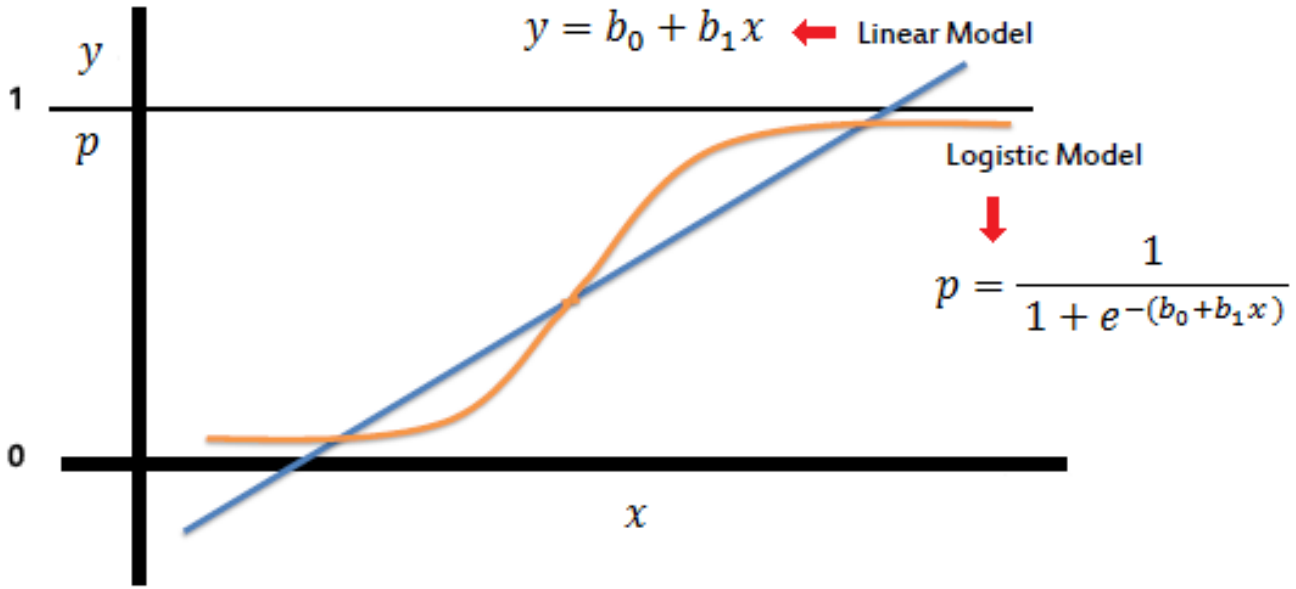

Virus
Similar
Human
Similar
Similar
Human
Human
Interact
Interact
Interact
Human
Human
Human
Human
Interact
Interact
Interact
Human
Human
Human

Scale of Real Data



The Prediction Model

Logistic Regression is a ML model used to make <u>binary</u> predictions



Source: Saedsayad

1. Similarity of virus to human protein

2. Number of motif matches

3. Frequency of same protein interaction

4. Interaction score

5. 1 or 0 for training data

Results of Logistic Regression

After testing on 45211 data points of which 5289 (11.7%) were interactions

The model had **89.58%** accuracy when run on the testing data after being given a set of training data

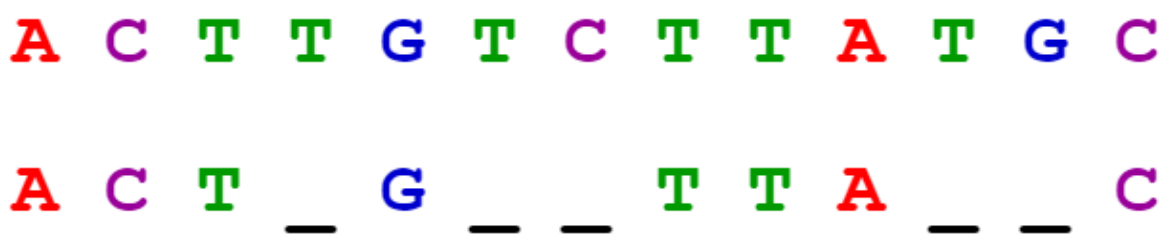
NOT INFORMATIVE

New Approach: Protein Sequence Alignment

Goal: Align amino acid chains to measure protein similarity in terms of interaction with other proteins

Default Algorithm: Needleman Wunsch (Dynamic Programming on amino acid sequence)

Problems: Cannot incorporate global structure



Source: Towardsdatascience

Global Alignment of Sequences

Needleman-Wunsch algorithm: dynamic programming to align two strings

Includes an affine gap penalty

A C T T G T C T T A T G C

A C T _ G _ _ T T A _ _ C

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Source: Wikipedia

Necessity of Structural Information

Physical structure of protein is known to have a strong correlation to the interaction type of the protein

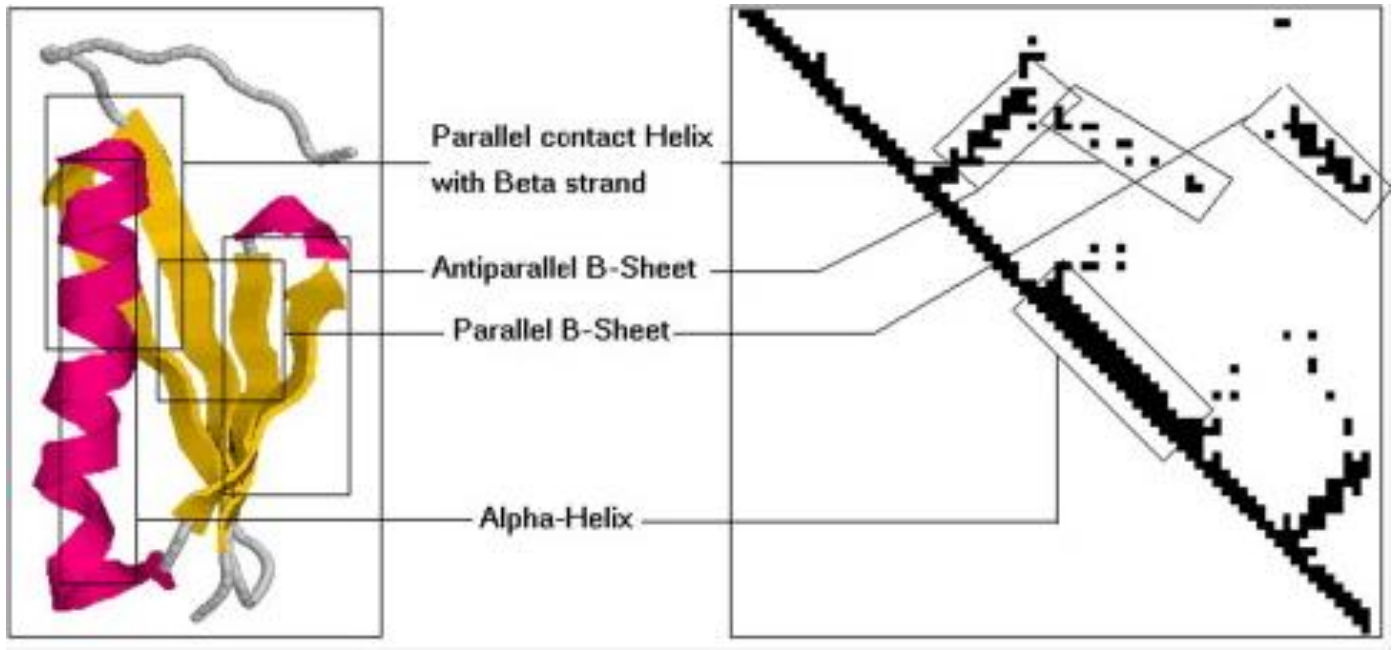
We want to be able to score proteins based on how similar their interactions are

Structural Information from Contact Maps

Proteins can be represented by their **contact map**: a binary matrix showing the presence of a contact between two residues

A **contact** is defined as a distance of less that 10 Å

The row and columns of the matrix represent the **residue chain** of the protein



Source: Science Direct

Our new approach

1. Represent contact maps as graphs

2. Structural similarities between vertices

3. Amino acid similarity

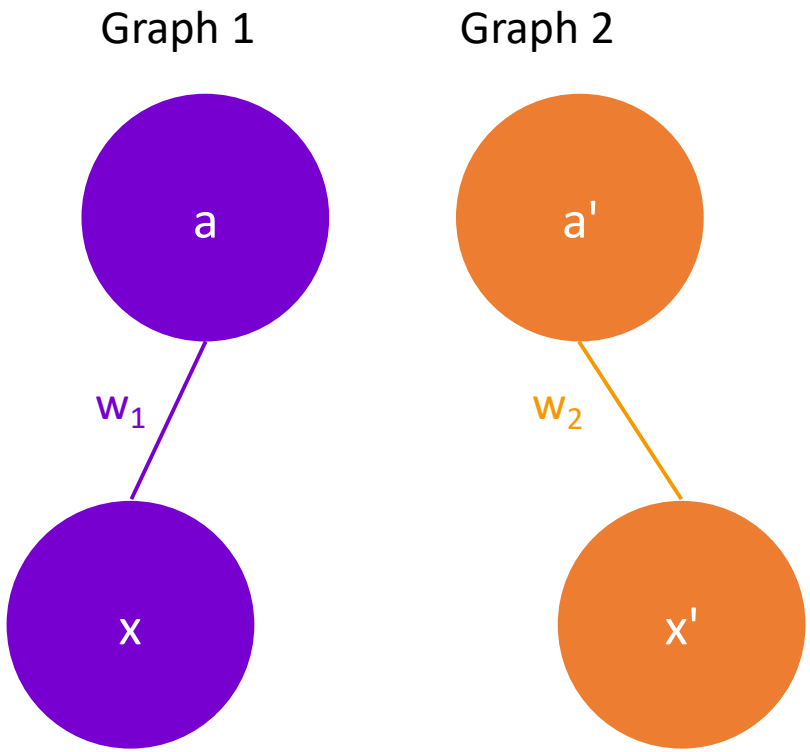
4. Combined payoff matrix

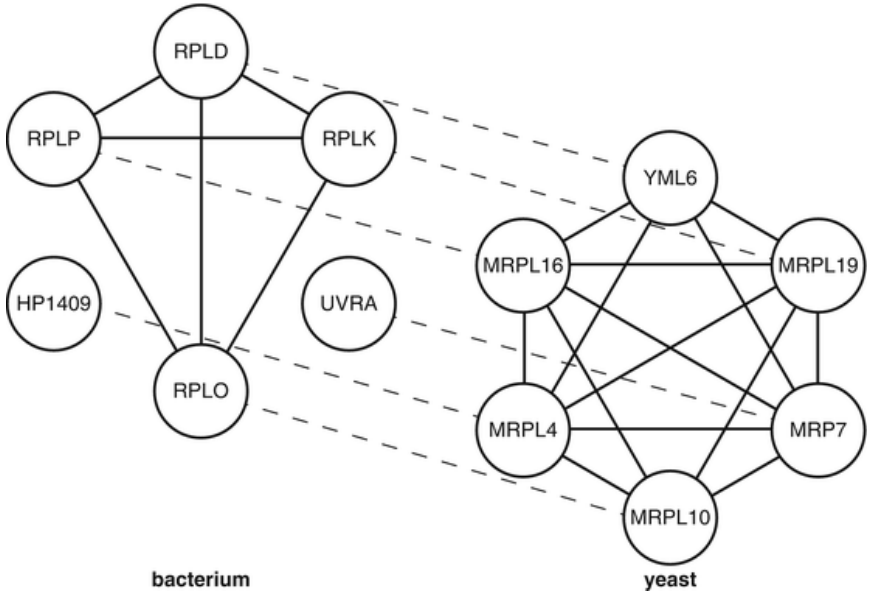
5. Sequence alignment

Graph Alignment

Weighted graphs are aligned by mapping each node to either an empty space or a node on another graph

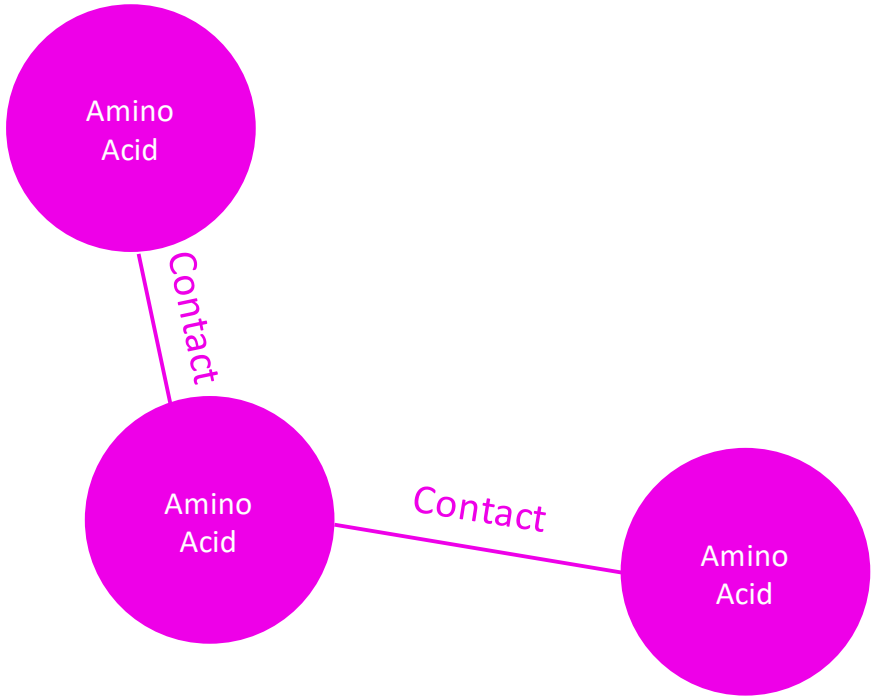
Goal: edges overlap as much as possible





Source: Graph Alignment, Protein Interaction Networks

Model Contact Map as Graph

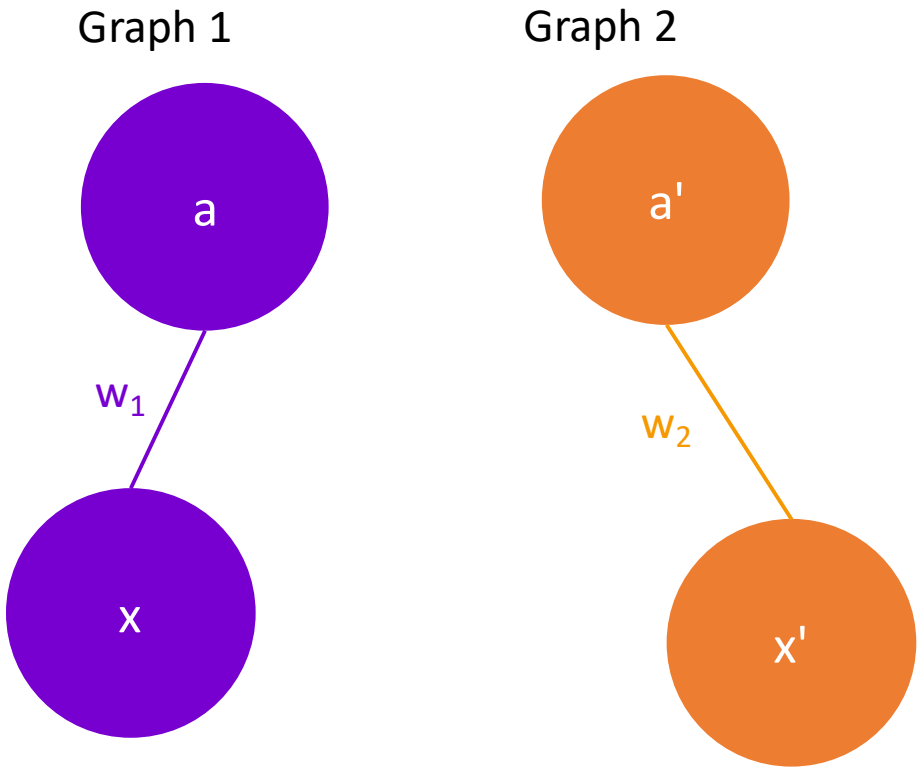


Graph Representation:
Vertices are amino acids
Edges are contacts

Goal: Combine graph alignment with sequence alignment

Idea: Compare graph structure without explicit graph alignment

Structural Similarity



Intuition: If a is mapped to a', then x is more likely to be mapped to x'

$R_{a,a'}$ is score for how likely we should map a to a'

Isorank Graph Alignment Algorithm

G,H graphs

$G_{i,j}$ represents the weight of the edge between nodes i,j

Matrix R:

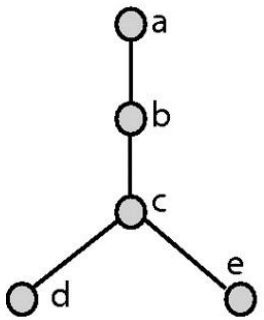
$$R_{a,a'} = \sum_{x \in G} \sum_{x' \in H} R_{x,x'} \frac{G_{a,x} H_{a',x'}}{deg(x) deg(x')}$$

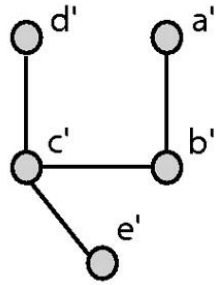
Where a,a' are vertices in G,H respectively

Calculating R

If reshape R to be a vector:

$$R = AR$$





R

	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

$R_{aa'} = \frac{1}{4} R_{bb'}$

$R_{bb'} = \frac{1}{3} R_{ac'} + \frac{1}{3} R_{a'c} + R_{aa'} + \frac{1}{9} R_{cc'}$

$R_{dd'} = \frac{1}{9} R_{cc'}$

$R_{cc'} = \frac{1}{4} R_{bb'} + \frac{1}{2} R_{be'} + \frac{1}{2} R_{bd'} + \frac{1}{2} R_{eb'} + \frac{1}{2} R_{db'} + R_{ee'} + R_{ed'} + R_{de'} + R_{dd'}$

Source: Global alignment of multiple protein interaction networks with application to functional orthology detection

$$A_{(a,a'),(x,x')} = \frac{G_{a,x} H_{a',x'}}{deg(x) deg(x')}$$

Can solve for R

Sequence Alignment

Blosum Substitution Matrix represents "similarity" between amino acids

Combine R matrix and Blosum to form payoff matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

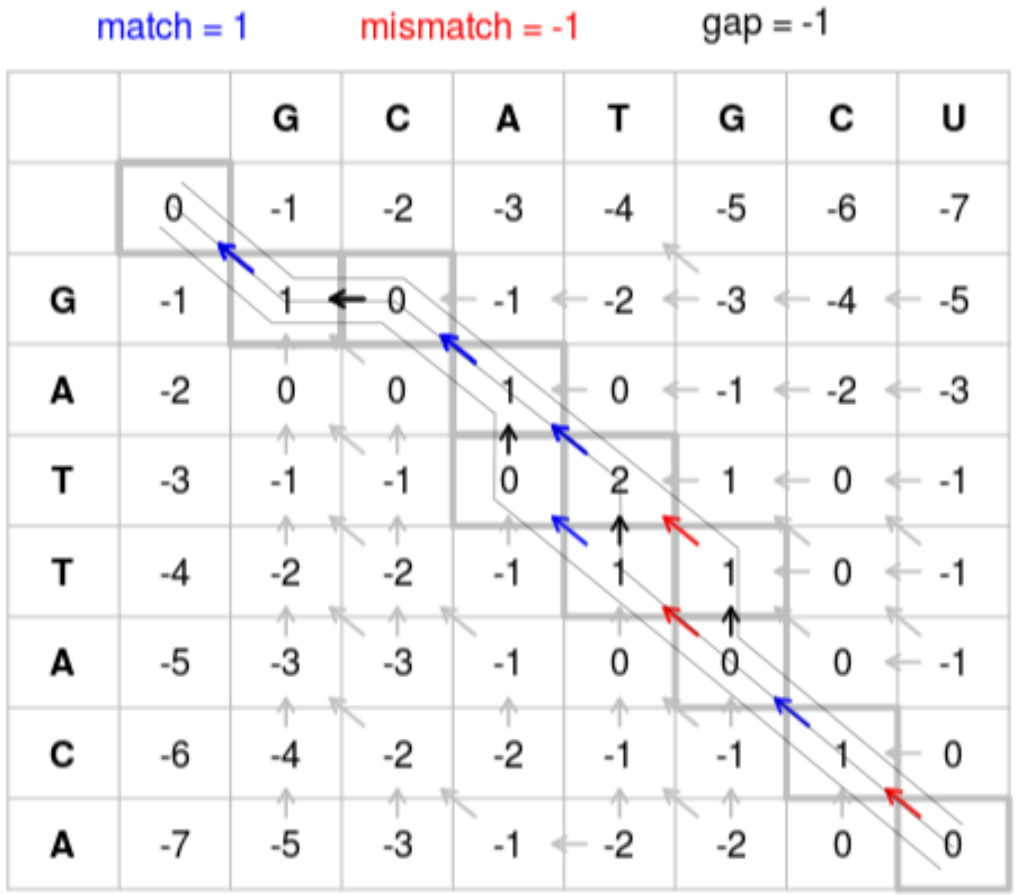
Source: Wikipedia

$$R'_{x,x'} = R_{x,x'} + \mathrm{BLOSUM}_{G_x,G_{x'}}$$

Needleman-Wunsch revisited

Needleman-Wunsch using R' as the payoff matrix

R' incorporates amino acid information and global protein structure



		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

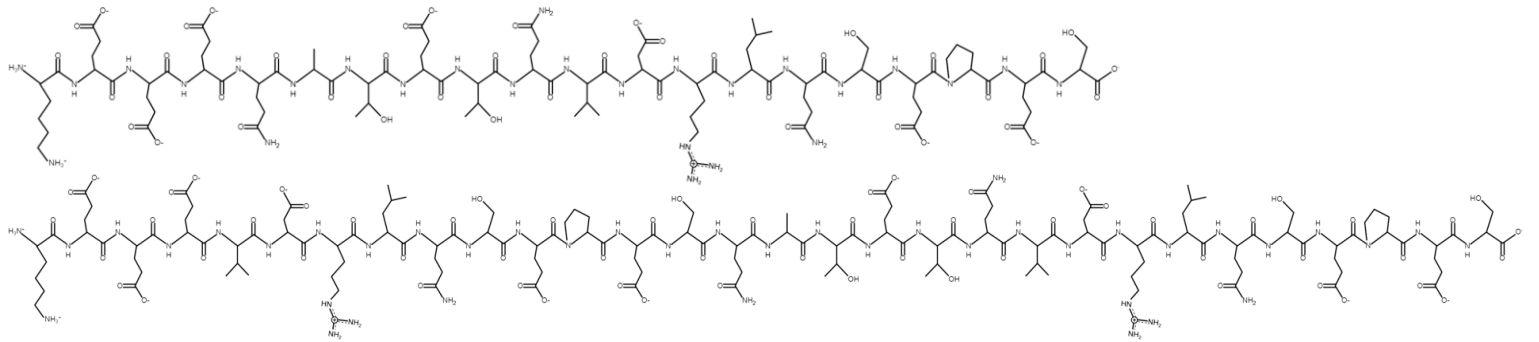
Example Result

Input1:

KEEEQATETQVDRLQSEPES

Input2:

KEEE**VDRLQSEPES**QATETQVDRLQSEPES



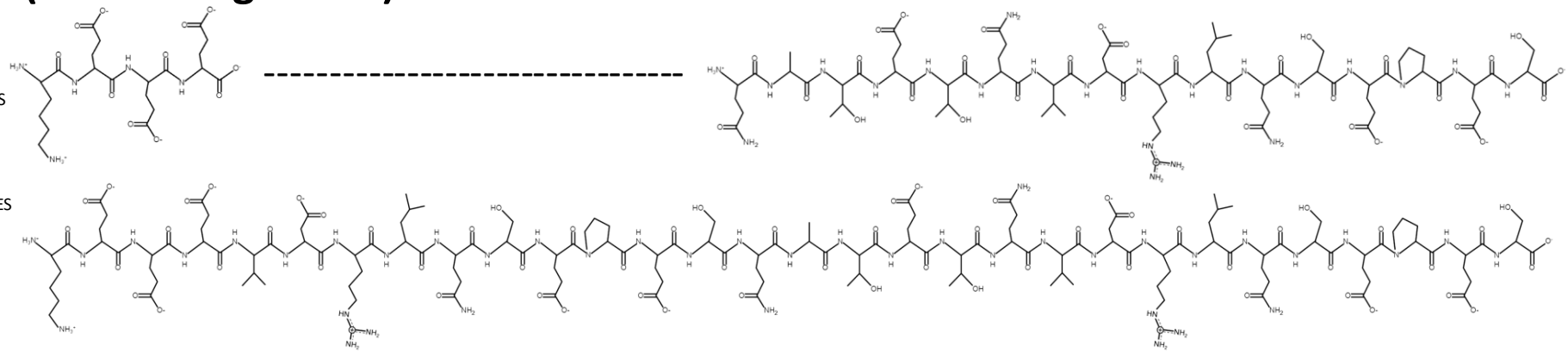
Our Algorithm: (Correct alignment)

Output1:

KEEE----------QATETQVDRLQSEPES

Output2:

KEEE**VDRLQSEPES**QATETQVDRLQSEPES



Note: BLOSUM only algorithm does not output correct alignment

Conclusion

We alignment the original contact map of the protein to an altered matrix where 10 residues were added to the chain

Replace the similarity scores in the logistic regression method with these scores

Algorithm is slow $O(n^4)$ time

Test our model on protein interface alignments

Thank you

- Younhun Kim, for being our mentor
- Upasana Das Adhikari, for suggesting the original project
- Dr. Slava Gerovitch and the PRIMES program, for giving us this opportunity

References

Rohit Singh, Jinbo Xu, Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proceedings of the National Academy of Sciences Sep 2008, 105 (35) 12763-12768; DOI:10.1073/pnas.0806627105

Needleman, Saul B. & Wunsch, Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. March 1970 48 (3): 443–53. DOI:10.1016/0022-2836(70)90057-4.