**Multimodal Data Fusion and Machine Learning for Deciphering Protein-Protein Interactions**

**Arghamitra Talukder**

**Concept of Operations**

Concept of Operations

for

Multimodal Data Fusion and Machine Learning for Deciphering Protein-Protein Interactions

Team <66>

Approved by:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Arghamitra Talukder 09/06/2020

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

John Lusher II, P.E. Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

T/A Date

**Change Record**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rev.** | **Date** | **Originator** | **Approvals** | **Description** |
| **1.0** | 09/06/2020 | Arghamitra |  | Concept of operation |

Table of Contents

[Table of Contents III](#_Toc50163898)

[List of Table IV](#_Toc50163899)

[List of Figures V](#_Toc50163900)

[1. Executive Summary 1](#_Toc50163901)

[2. Introduction 2](#_Toc50163902)

[2.1. Background 2](#_Toc50163903)

[2.2. Overview 2](#_Toc50163904)

[2.3. Referenced Documents and Standards 3](#_Toc50163905)

[3. Operating Concept 5](#_Toc50163906)

[3.1. Scope 5](#_Toc50163907)

[3.2. Operational Description and Constraints 5](#_Toc50163908)

[3.3. System Description 5](#_Toc50163909)

[3.4. Modes of Operations 6](#_Toc50163910)

[3.5. Users 6](#_Toc50163911)

[3.6. Support 7](#_Toc50163912)

[4. Scenario(s) 8](#_Toc50163913)

[4.1. Interprotein interaction 8](#_Toc50163914)

[4.2. Intra protein interction 8](#_Toc50163915)

[5. Analysis 8](#_Toc50163916)

[5.1. Summary of Proposed Improvements 8](#_Toc50163917)

[5.2. Disadvantages and Limitations 8](#_Toc50163918)

[5.3. Alternatives 8](#_Toc50163919)

List of Figures

[Figure 1: The proposed protein predicting model overview 3](#_Toc50164401)

[Figure 2: The system description 6](#_Toc50164402)

List of Tables

[Table 1: Constraints and Used Media 5](#_Toc50164597)

# Executive Summary

Protein-protein interaction is one of the most important biological process and it helps to decode a lot of Molecular Biology questions. Given the importance of PPI their interaction in 3-dimensional space must be studied thoroughly. Due to the huge quantity of unknown biological interactions, it is a next to impossible task to identify each physical protein interaction in person; not only because the identification presents an enormous quantity but also it is a costly and resource intensive job. Though from time to time various experimental and computational methods have been applied to predict the PPI, a knowledge gap is there to understand their 3-dimensional interactions. This research project aims to use the existing data and available tools of machine learning resulting in an algorithm to predict protein-protein interaction. The algorithm will use data science to match the existing patterns with a model in the light of physics and biology. The success matric of the algorithm will be the accuracy of the testing and validation group of data; it also aims to cover a broad range of scope making it more versatile.

# Introduction

A functional human body is made of a lot of active organs, different macro and micro molecules. One of the most important cellular molecules is protein. Protein contributes in most biological processes including genetic expression, intercellular communication, morphology, nutrition absorption and so on (Thomas et al). Proteins are made of a character string where the characters represent amino acid. The amino acids bond together in different configurations to express the specific functionalities as proteins. As the mechanisms of the human body are unrevealed, one context was very clear that most of the proteins interact with each other and to understand their behaviors they should be analyzed from the perspective of protein-protein interaction. There are experimental methods as well as computational methods. Given the short scale application of the experimental techniques like affinity purification, yeast two hybrid, co-immunoprecipitation, computational methods are more suitable approaches to follow.

## Background

The computational method adopted homology-based approaches like interolog search. Interolog search is based on the principle that interactions are conserved and interlogs are homologous pairs of protein interactions across different species. The homology-based method also includes phylogenetic similarities which relates to the common ancestor proteins among species (Abbasi et al). The simulation-based methods include protein docking. Protein docking is molecular modeling which predicts the mutual orientation (Tradigo et al). A lot of machine learning techniques have been also applied based on protein sequence, structure and function. The limitations with these approaches are the difficulties to model any conformational changes and lack of thorough understanding of the binding mechanism (Abbasi et al).

This project offers a computational method to predict the overview how the proteins interact in 3-dimensional space. An accurate PPI prediction model will serve several objectives including pathways for unknown proteins, different binding modes, specificity of protein based multiple targets, effectiveness of drugs, design of new protein etc. The goal of this project is to develop an algorithm using machine learning to predict the physical protein-protein interactions expressing them in different modalities like text, graph and image

## Overview

Given two sets of proteins, the system will be designed to predict the interaction probability between two amino acids between the protein. In the proposed algorithm the linear sequence (text), inter protein contact map (graph) and the 3-dimensional protein interaction(image) will integrated in a contact predictor. The entire mode of operation can be summarized in the following figure.

|  |  |  |
| --- | --- | --- |
| A close up of a logo  Description automatically generated | Input of two different protein | Output:  Do they have contact or no contact  A picture containing food, fruit  Description automatically generated |

Figure 1: The proposed protein predicting model overview

The figure is reprinted from Nooren

## Referenced Documents and Standards

Tradigo, Giuseppe, et al. “Algorithms for Structure Comparison and Analysis: Docking.” *Encyclopedia of Bioinformatics and Computational Biology*, Academic Press, 6 Sept. 2018, [www.sciencedirect.com/science/article/pii/B9780128096338204858](http://www.sciencedirect.com/science/article/pii/B9780128096338204858).

Abbasi, Wajid Arshad & Minhas, Fayyaz ul Amir Afsar. (2018). Problems in Protein-Protein Interactions (A Literature Review).

Thomas, Neil, et al. “Can We Learn the Language of Proteins?” *The Berkeley Artificial Intelligence Research Blog*, 4 Nov. 2019, bair.berkeley.edu/blog/2019/11/04/proteins/.

Muscat, M., Croce, G., Sarti, E., & Weigt, M. (n.d.). FilterDCA: Interpretable supervised contact prediction using inter-domain coevolution. doi:https://doi.org/10.1101/2019.12.24.887877

Protein structure: Primary, secondary, tertiary & quatrenary (article). (n.d.). Retrieved September 05, 2020, from https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

Bittrich, S., Schroeder, M., & Labudde, D. (2019, December 06). StructureDistiller: Structural relevance scoring identifies the most informative entries of a contact map. Retrieved September 05, 2020, from https://www.nature.com/articles/s41598-019-55047-4

Nooren, I., & Thornton, J. (2003, July 15). Diversity of protein–protein interactions. Retrieved September 05, 2020, from https://www.embopress.org/doi/10.1093/emboj/cdg35

# Operating Concept

## Scope

The project has a very broad and important scopes to understand human physiology and disease protection and prevention. This algorithm can be used to manufacture and design drugs and artificial protein which is a current issue in this COVID situation. With this kind of precise ML tools can give a very fast direction to vaccine development to create the corona virus antibody

With the use of this tool a lot more about human physiology can be known. Though the medicine and genetic engineering have come a long way, still there are a lot of physiological mysteries which are yet to be solved like the cause and prevention of Alzheimer’s or schizophrenia. This tool may provide a new perspective to solve this problem by predicting a precise physical bonding between two unknown protein.

## Operational Description and Constraints

Already there are a lot of tools to take two different input and make a multi sequence alignment (MSA). With two different protein the system will take the MSA of two unknown proteins. With different features from text, graph and image of existing protein structures the ML model will be trained and will predict if the unknown proteins have interaction or no interaction.

Among other constraints for the database for now PDB will be used and Google drive and git will be used as storage platforms. As IDE PyCharm, Google Colab, Jupiter notebook will be used. The Python ML tools will be used to analyze the data and train, test the model

|  |  |
| --- | --- |
| **Constrains** | **Media use** |
| Operating system | Linux, Windows |
| Storage | GitHub, Google Drive |
| IDE | PyCharm, Google Colab, Jupiter Notebook |
| ML tools | Python in built functions, specifically pyDCA |

Table 1: Constraints and Used Media

## System Description

The proteins will be used in three modalities: text, graph and image. When the protein is represented as texts, that means only the amino acid sequence of the protein has been taken into consideration. There is a very popular method has been developed to predict intra protein interaction which is called Direct Coupling Analysis. This method takes consideration the text or amino acid sequence of the protein and extracts the coevolutionary features from them.

From the graph the contact map information is extracted. From the image the residue contacts have been used as features.

The entire system can be summarized in the following figure:

A close up of a screen

Description automatically generated

[1]

Inputs

Two unknown protein

Text

A picture containing drawing

Description automatically generated

[2]

Output

Features

Image

Graph

A picture containing map

Description automatically generated

[3]

Do they have contact or no contact

Figure 2: The system description.

The figure is reprinted from Muscat [1], Bittrich [2] and Protein structure: Primary, secondary, tertiary & quaternary [3]

## Modes of Operations

There can be main one mode of operation which is taking the input of MSA the ML model will predict the output. There is a possibility to add another mode which takes just the name of the protein and has the MSA operation integrated in it.

## Users

This model can be used by other research groups and medical teams

## Support

The goal of this project is to publish a paper. Thus, elaborate documentation, calculation, reference will be provided.

# Scenarios

## Interprotein interaction

This algorithm and the ML model will be used to predict the interprotein interaction

## Intra protein and RNA Sequence Analysis

In the previous researches, some of the algorithm like pyDCA has been used in RNA sequence analysis. Another paper on the method filterDCA has been used in intra protein residue contact prediction. This model also can be used in these cases with little modification

# Analysis

## Summary of Proposed Improvements

In summary, the main exclusivity of the proposed model lies into different modality representation. So far, CNN has been used taking the feature input from DCA score and the secondary structure. In the proposed model we will representing the protein into three different forms and extract the features specific to that form and use them as feature

## Disadvantages and Limitations

In any machine learning application, the computation power and time always has been an issue. Thus, the main limitation is the computation time to train and test the model

## Alternatives

One of the alternative solutions is to limit the dataset but the tradeoff is to overfit the model. The small amount of dataset also hampers the performance of the model.