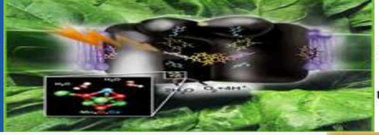
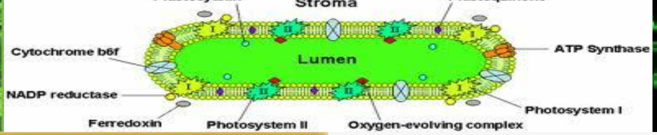
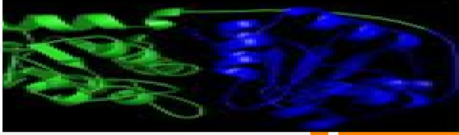






Advisor: Dr. Wei Chen

Senior Project II

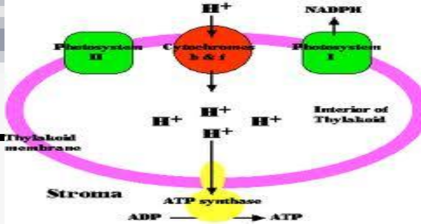
**Tennessee State University**
**College of Engineering**
**Department of Computer Science**

# DETECTION OF INTERACTION SITES OF PROTEINS

by
Pankaj Mishra and Anthony Burkeen

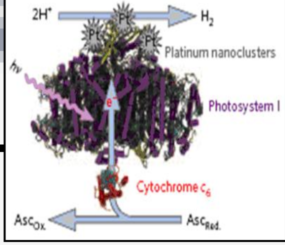
Funded by TN-Score and National Science Foundation

# Table of Contents

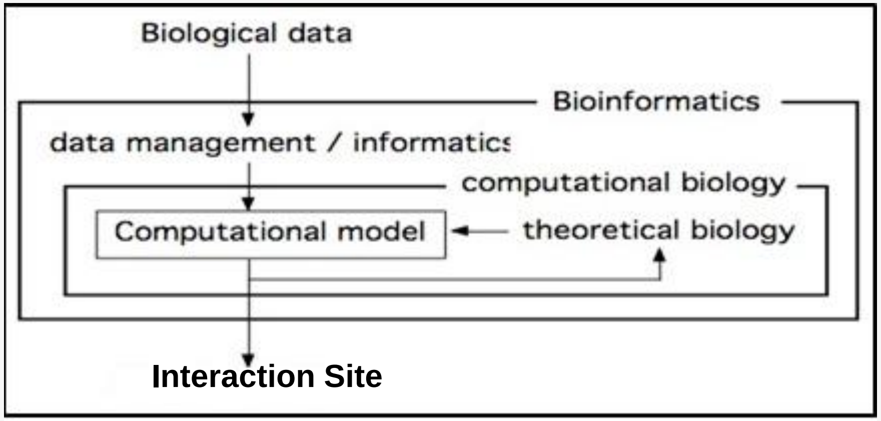


- Introduction
- Problem Statement
- Goal and Objectives
- Requirement Analysis
- System Design
  - System Model
  - Dataset preparation
  - Amino Acids bond Analysis
  - Protein interaction sites prediction algorithm
- Implementation
- Result
- Conclusion

# Introduction



- This project focuses on detecting the interaction sites of Photosystem I and Cytochrome protein family. The interaction of the protein pairs of these two families is used to speed up hydrogen producing in photosystem I.



Biological data

Bioinformatics

data management / informatics

computational biology

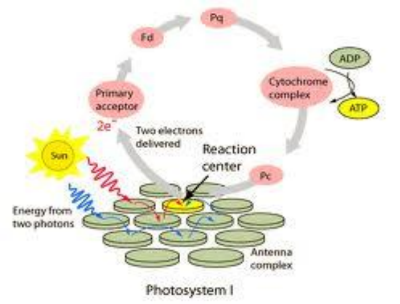
Computational model

theoretical biology

Interaction Site

- Computational approaches are proposed for predicting interaction sites of protein pairs of cytochrome c6 and photosystem I unit PsaF (photo system I family).

# Problem Statement



- Finding new energy sources such as “Hydrogen”.
- Utilizing protein interaction to efficiently produce hydrogen.
- Identify the best proteins that speed up the hydrogen producing.

# Mechanisms of Producing Hydrogen in Photosystem

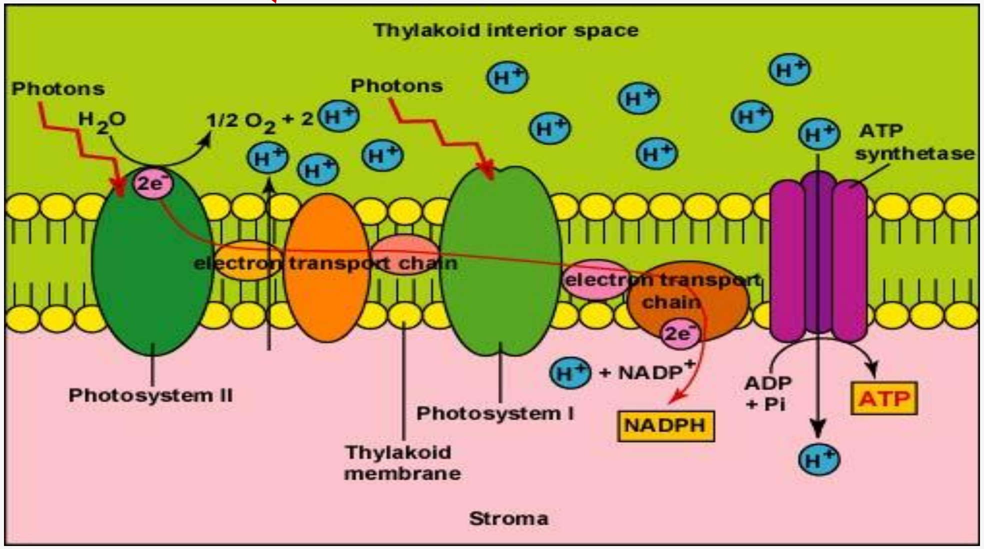
*Difference between natural process and System process*

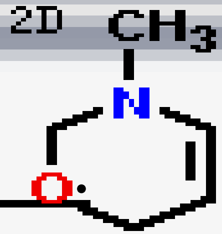
## Natural Process

- Slow process
- Not quantitative

## Artificial Process

- Fast process
- Quantitative





# Goal and Objectives

**Goal:**

Computationally detecting the interaction sites of proteins from PsaF family and Cytochrome family

**Objectives:**

- ❖ Find proper protein sequences from PsaF family and Cytochrome family in same organism.
- ❖ Identify bonding properties of amino acids
- ❖ Develop the interaction prediction algorithms
- ❖ Evaluate and analyze data

# Requirement Analysis

- The requirement analysis determines the requirements and conditions for the detection of the protein interaction sites.

- Requirements are divided into two parts as functional requirements and non-functional requirements.

# Functional Requirements

- Protein sequences are required for detecting the interaction sites.

- National Center for Biotechnology Information’s BLAST database and other databases are required for retrieving the protein sequences.

- The protein candidates have to be found from the interaction sites prediction algorithm.

- Database (NCBI, Gene Bank) are required for fundamental sequence analysis.

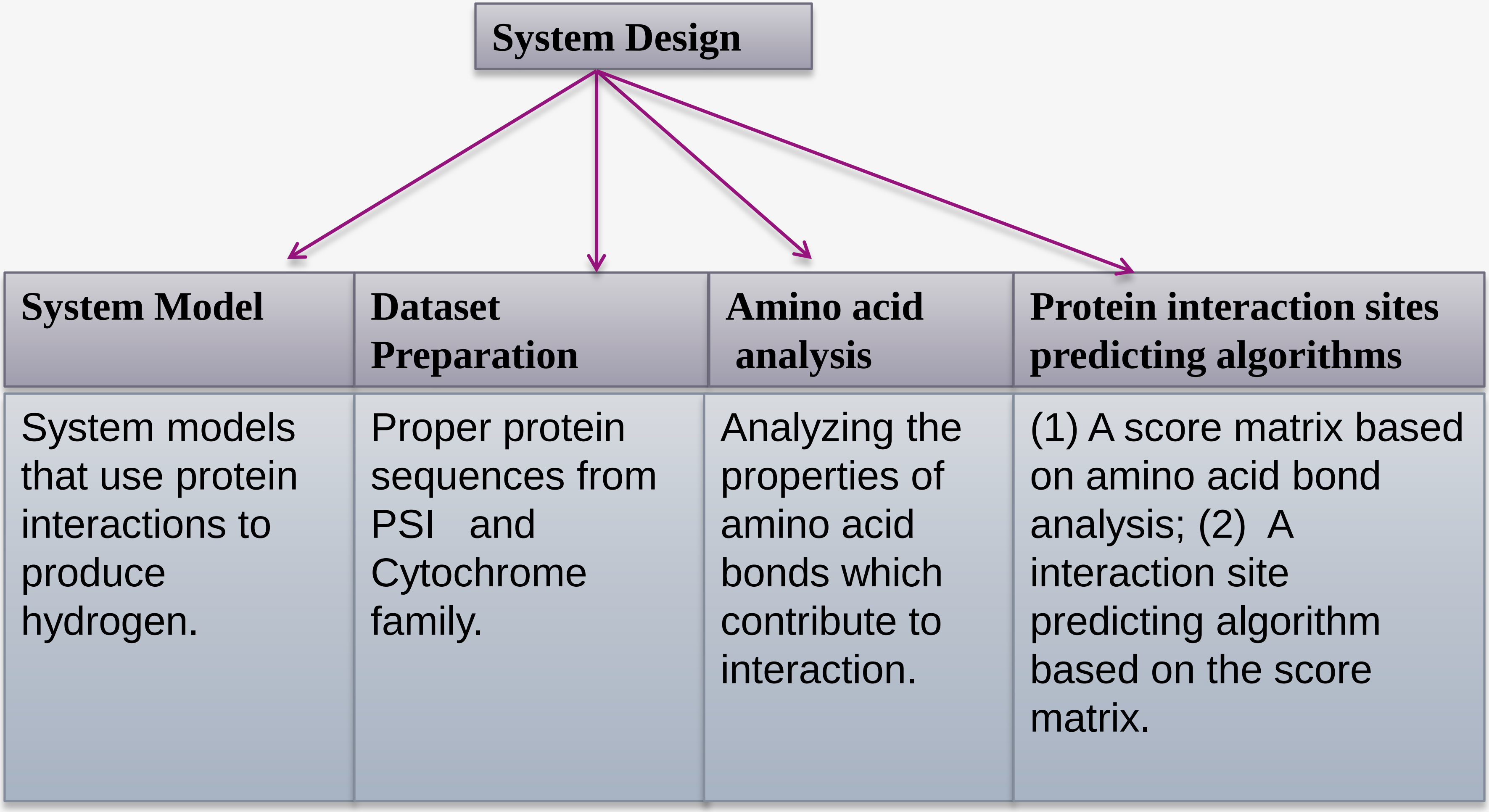
# Non- functional Requirements

- Results from smaller datasets should demonstrate the influence of interaction sites of proteins.

- 20 protein sequences (totally 86 pairs of proteins) are required from both PsaF family and C6 family.

- The proposed approach, algorithm and software must be implemented and evaluated.

# System Design

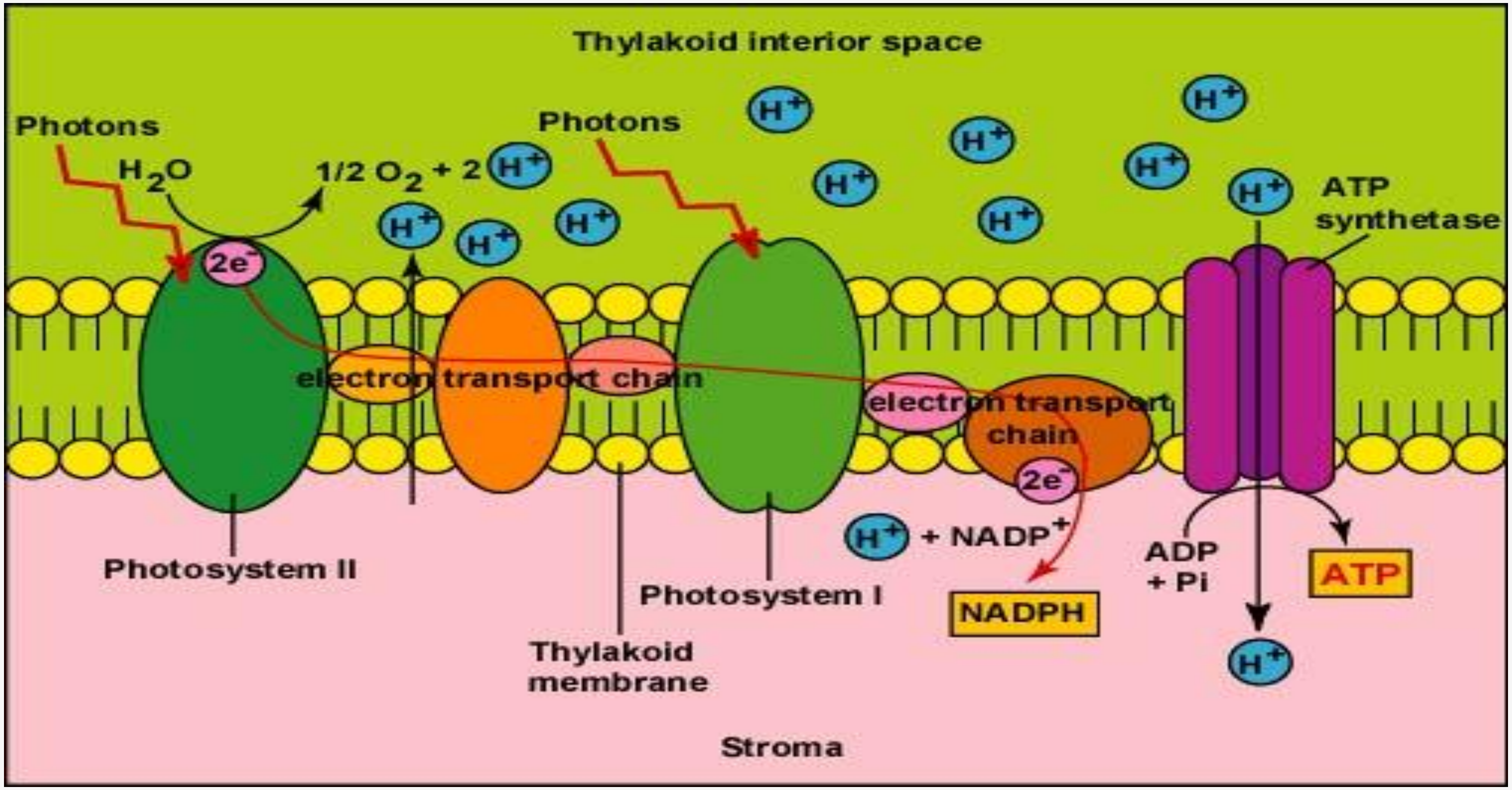


System Design

| System Model | Dataset Preparation | Amino acid analysis | Protein interaction sites predicting algorithms |
|---|---|---|---|
| System models that use protein interactions to produce hydrogen. | Proper protein sequences from PSI and Cytochrome family. | Analyzing the properties of amino acid bonds which contribute to interaction. | (1) A score matrix based on amino acid bond analysis; (2) A interaction site predicting algorithm based on the score matrix. |

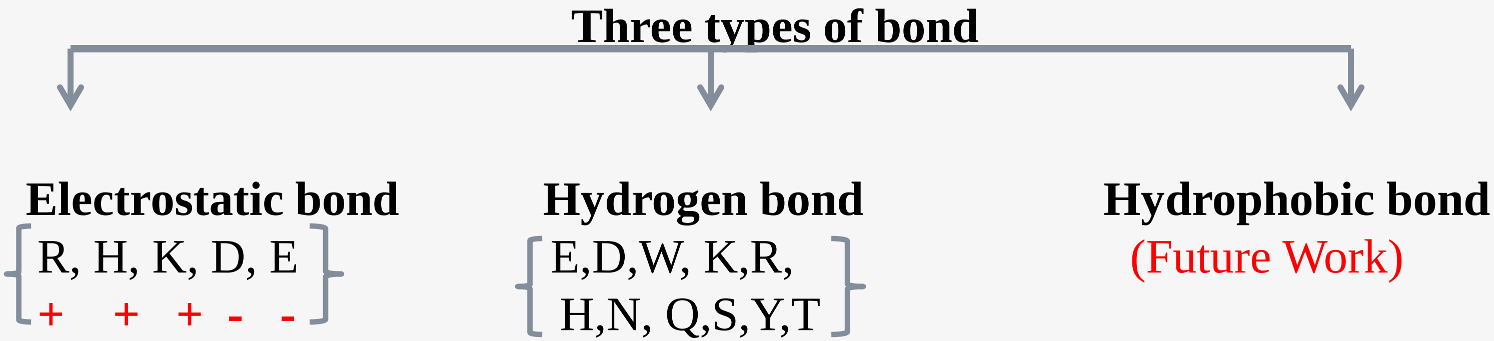
# System Design

## System Model



# System Design

## Amino Acid Bond Analysis



**Three types of bond**

**Electrostatic bond**

R, H, K, D, E
+ + + - -

**Hydrogen bond**

E,D,W, K,R,
H,N, Q,S,Y,T

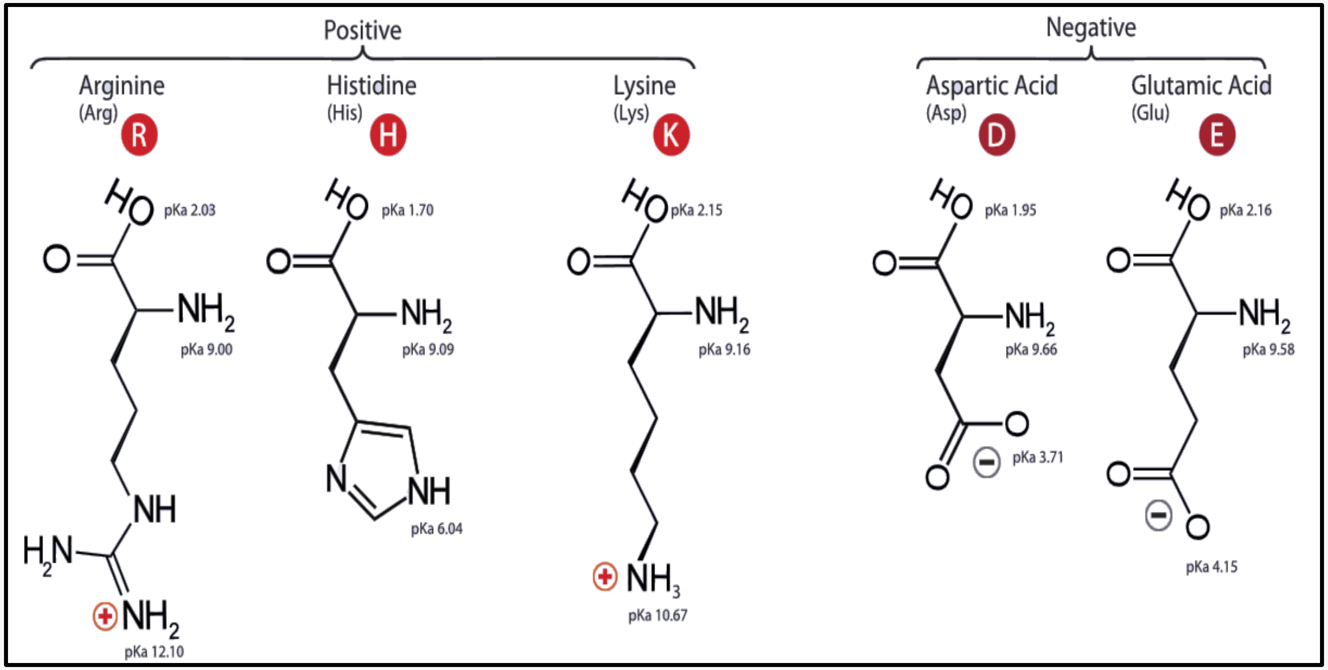
**Hydrophobic bond**

(Future Work)

**This research focuses on Electrostatic and hydrogen bond induced interaction.**

# System Design

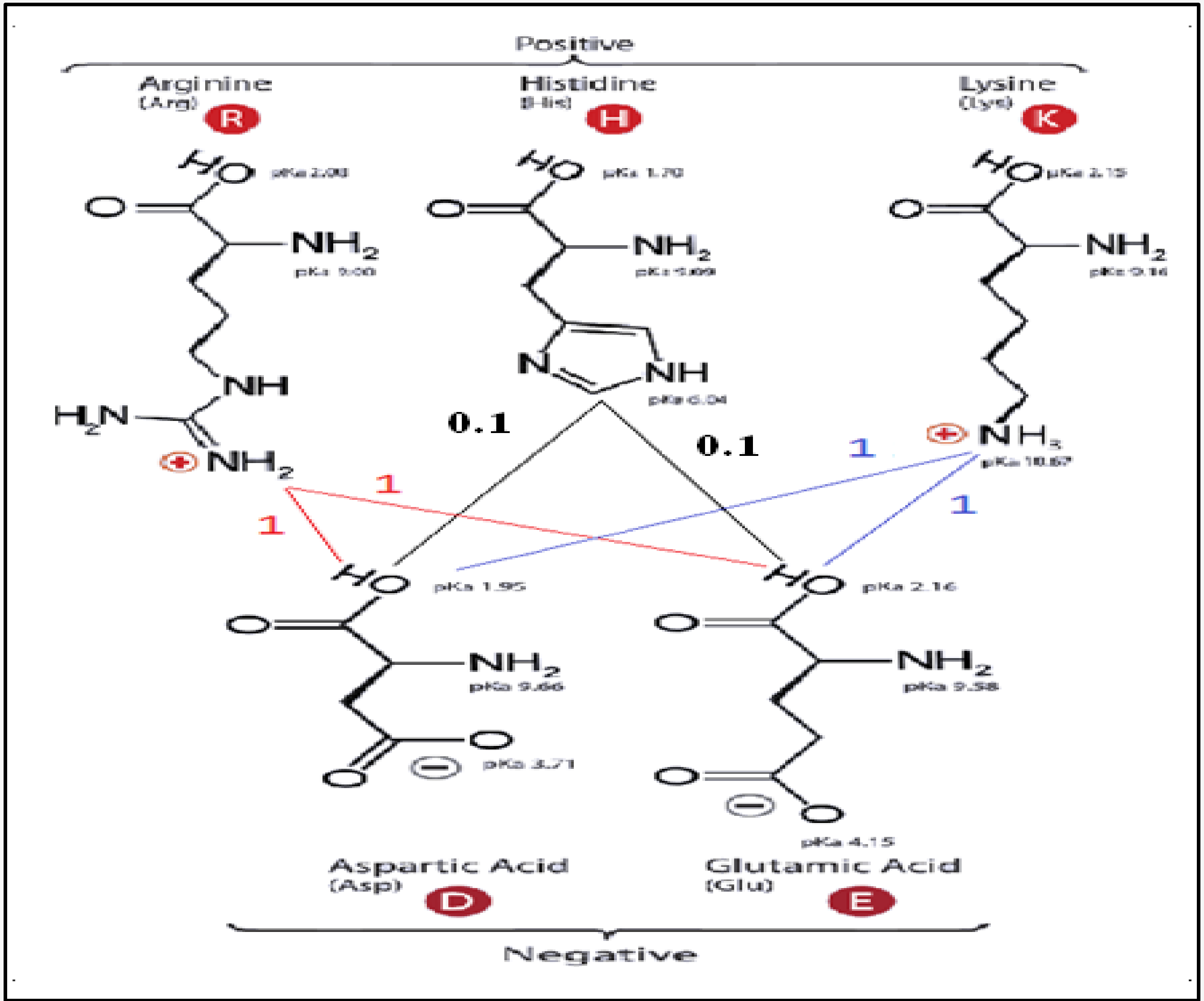
## Amino Acid Analysis (*Electrostatic bond* )



Argignine, Histidine, Lysine, Aspartic Acid, Glutamic Acid

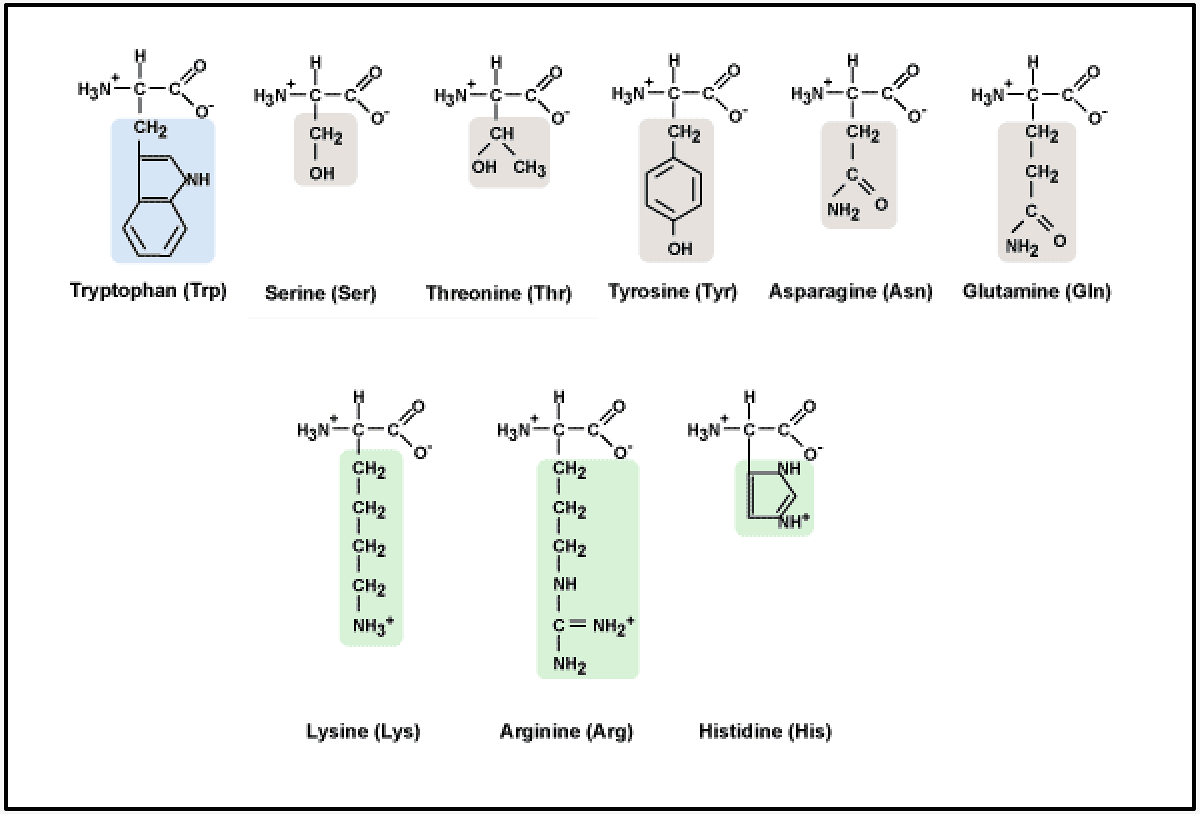
# System Design (*Electrostatic bond*)

**Score Matrix Design:** Bonding strengths between the four amino acids.



# System Design

## Amino Acid Analysis (*Hydrogen bond*)

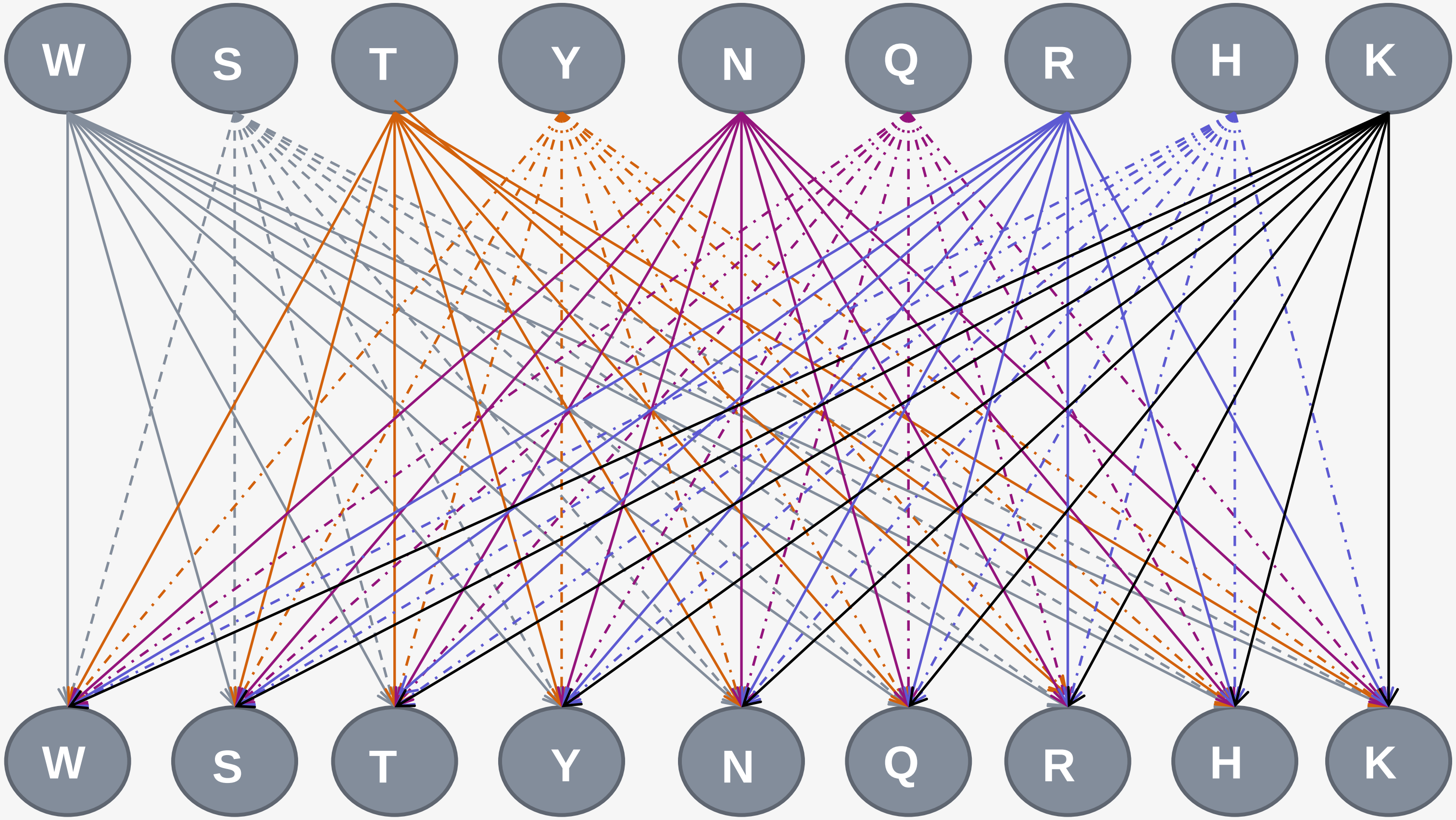


# System Design (Hydrogen bond + Electrostatic Bond)

## Protein Name

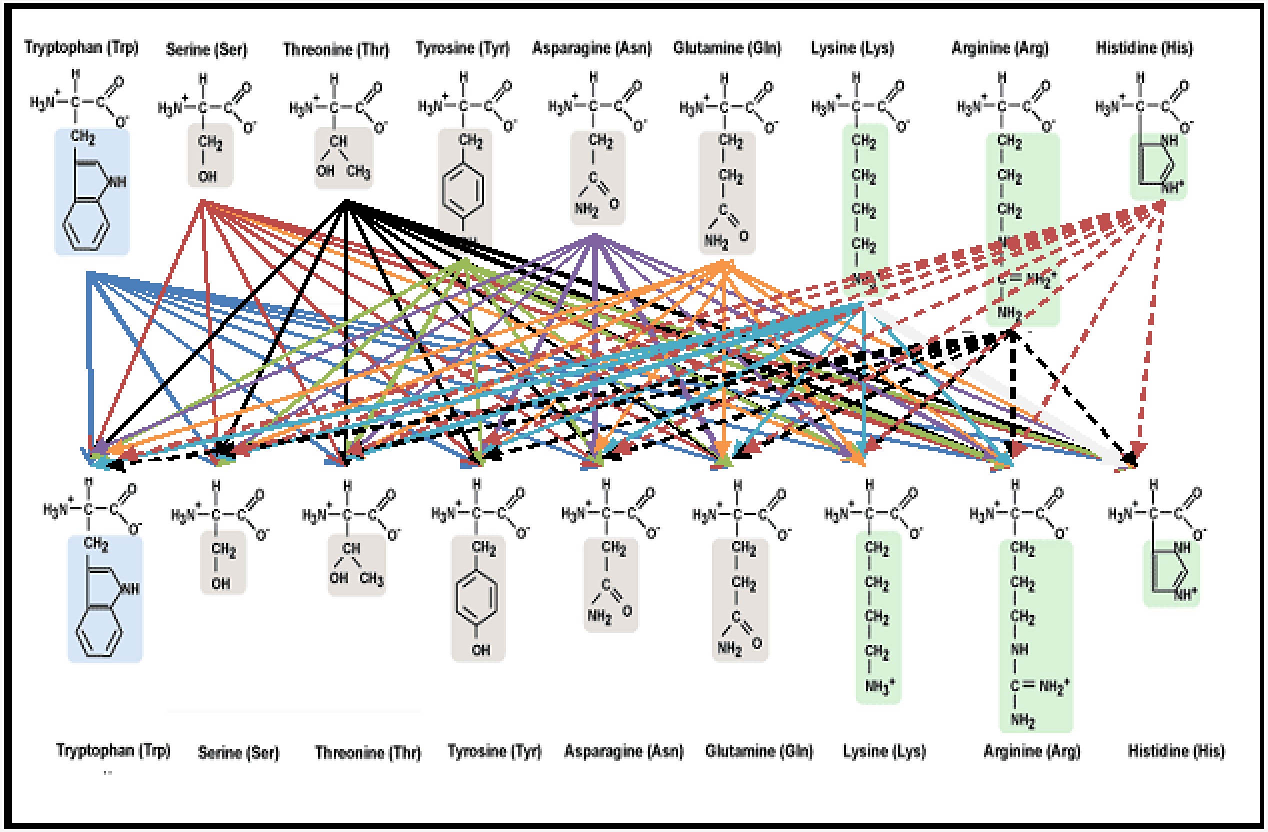
| | |
|---|---|
| W | Tyrptophan |
| S | Serine |
| T | Threonie |
| Y | Tyrosine |
| N | Asparagine |
| Q | Glutamine |
| R | Arginine |
| H | Histidine |
| K | Lysine |

# System Design (Hydrogen bond + Ionic Bond)



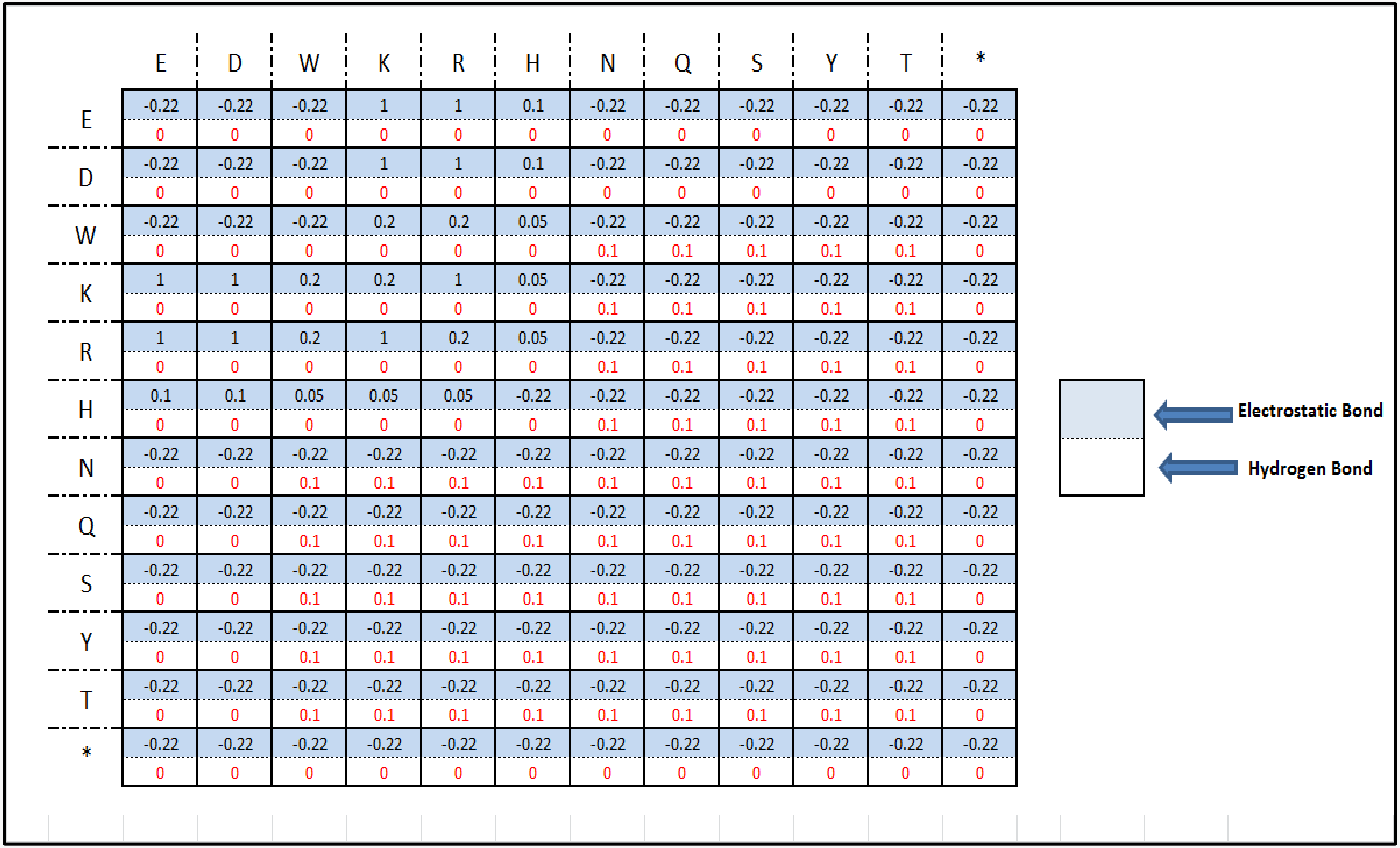
# System Design (Hydrogen bond + Ionic Bond)

Score Matrix Design: Bonding strengths between the nine amino acids.



# System Design

Weight of amino acids in hydrogen and electrostatic bond



| | E | D | W | K | R | H | N | Q | S | Y | T | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | 1<br>0 | 1<br>0 | 0.1<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 |
| D | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | 1<br>0 | 1<br>0 | 0.1<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 |
| W | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | 0.2<br>0 | 0.2<br>0 | 0.05<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| K | 1<br>0 | 1<br>0 | 0.2<br>0 | 0.2<br>0 | 1<br>0 | 0.05<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| R | 1<br>0 | 1<br>0 | 0.2<br>0 | 1<br>0 | 0.2<br>0 | 0.05<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| H | 0.1<br>0 | 0.1<br>0 | 0.05<br>0 | 0.05<br>0 | 0.05<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| N | -0.22<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| Q | -0.22<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| S | -0.22<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| Y | -0.22<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| T | -0.22<br>0 | -0.22<br>0 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0.1 | -0.22<br>0 |
| * | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 | -0.22<br>0 |

Electrostatic Bond

Hydrogen Bond

# System Design

## Protein interaction sites predicting algorithm

Given a pair of cyt c6 and PsaF protein sequences.

Step 1 (initialization): 0 in the first row of A from A[0,1], …, A[0,n]. 0 in the first column of A from A[1,0], …, A[m,0].

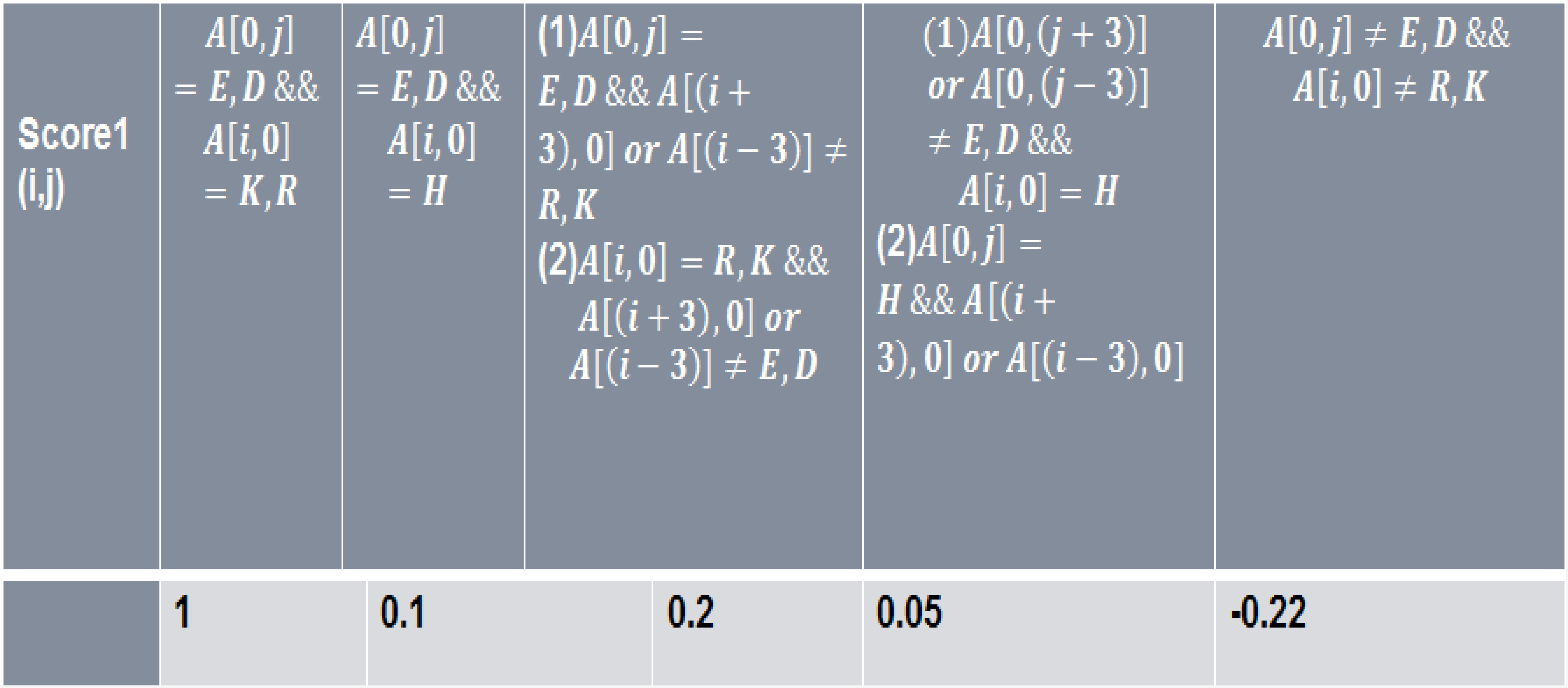
Step 2  A[i,j] = A[i-1,j-1]+max{score(i,j), 0}, where score(i,j) is given in the following table:

```
PredicUsingWindow(X,Y,W)
Step 1: Calculate matrix $S$
      for $i = 0$ to $m$ do $S[i,0] = 0$;
      for $j = 0$ to $n$ do $S[0,j] = 0$;
    for $i = 1$ to $m$
      for $j =1$ to $n$
        $S[i,j] = \max\{S[i,j]+W(i,j), 0\}$;
```

Step 3 Select the highest k scores from A.

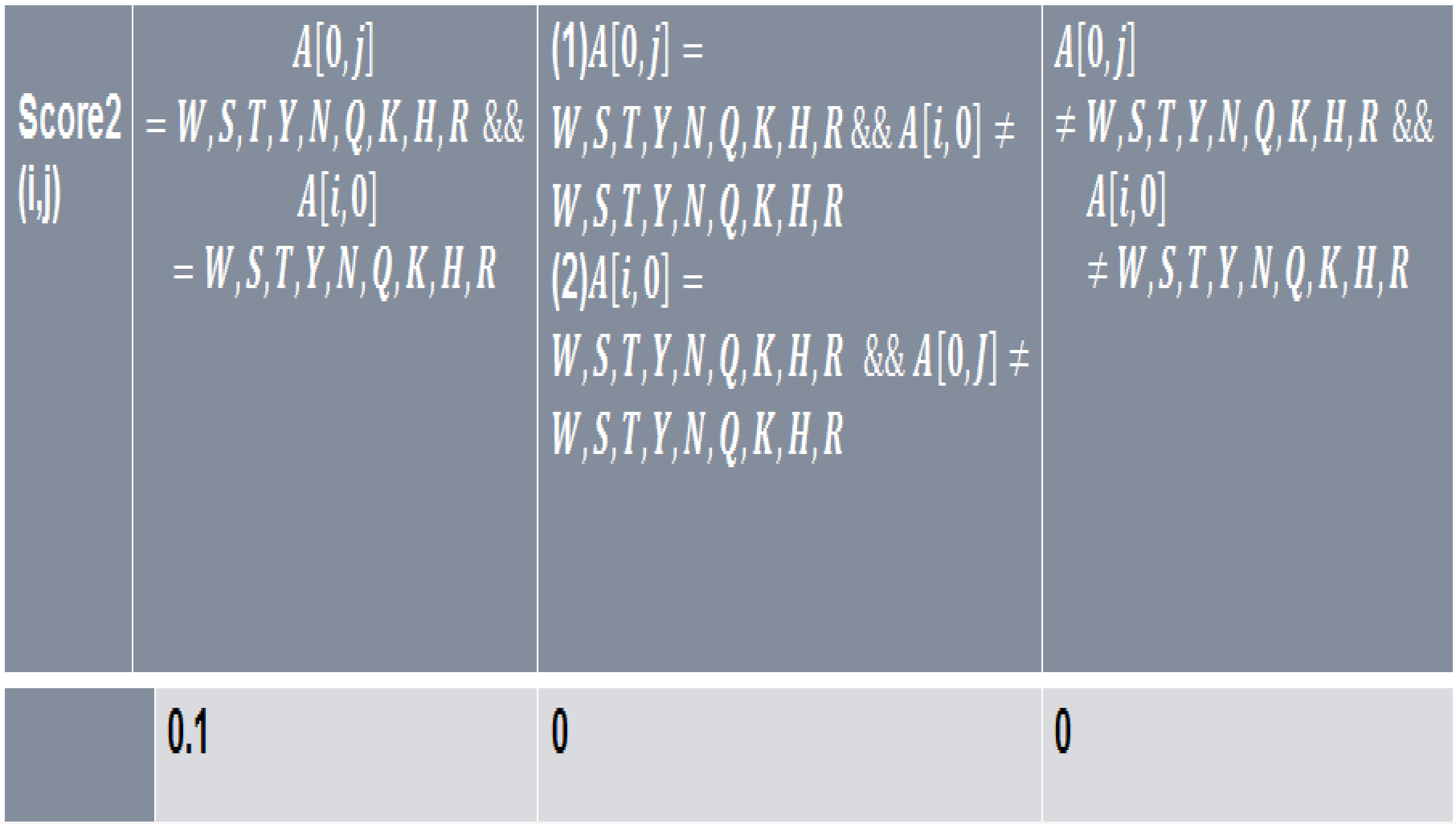
Step 4 Track back in A to get k interaction sites.

# Matrixes Scoring for Electrostatic Bond



| Score1 (i,j) | $A[0,j] = E,D \&\& A[i,0] = K,R$ | $A[0,j] = E,D \&\& A[i,0] = H$ | (1)$A[0,j] = E,D \&\& A[(i+3),0]$ $or$ $A[(i-3)] \neq R,K$<br>(2)$A[i,0] = R,K \&\& A[(i+3),0]$ $or$ $A[(i-3)] \neq E,D$ | (1)$A[0,(j+3)]$ $or$ $A[0,(j-3)] \neq E,D \&\& A[i,0] = H$<br>(2)$A[0,j] = H \&\& A[(i+3),0]$ $or$ $A[(i-3),0]$ | $A[0,j] \neq E,D \&\& A[i,0] \neq R,K$ |
|---|---|---|---|---|---|
|  | 1 | 0.1 | 0.2 | 0.05 | -0.22 |

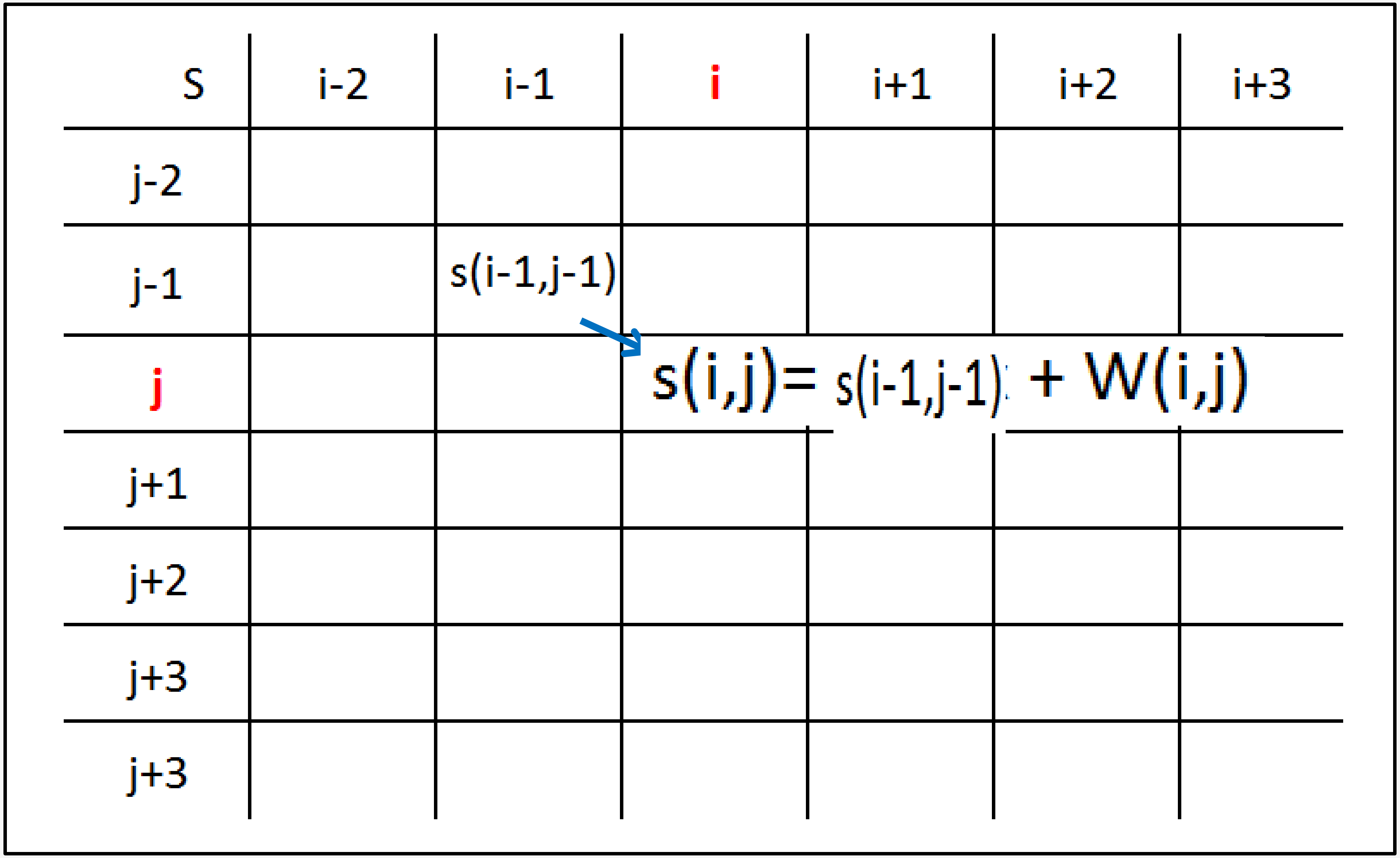
# Matrixes Scoring for Hydrogen Bond



| Score2 (i,j) | $A[0,j] = W,S,T,Y,N,Q,K,H,R$ && $A[i,0] = W,S,T,Y,N,Q,K,H,R$ | (1)$A[0,j] = W,S,T,Y,N,Q,K,H,R$ && $A[i,0] \neq W,S,T,Y,N,Q,K,H,R$<br>(2)$A[i,0] = W,S,T,Y,N,Q,K,H,R$ && $A[0,J] \neq W,S,T,Y,N,Q,K,H,R$ | $A[0,j] \neq W,S,T,Y,N,Q,K,H,R$ && $A[i,0] \neq W,S,T,Y,N,Q,K,H,R$ |
|---|---|---|---|
|  | 0.1 | 0 | 0 |

# Algorithm uses the weight scheme with window

The score at s(i, j) is decided by the score at s(i-1,j-1) and weight W(i,j)



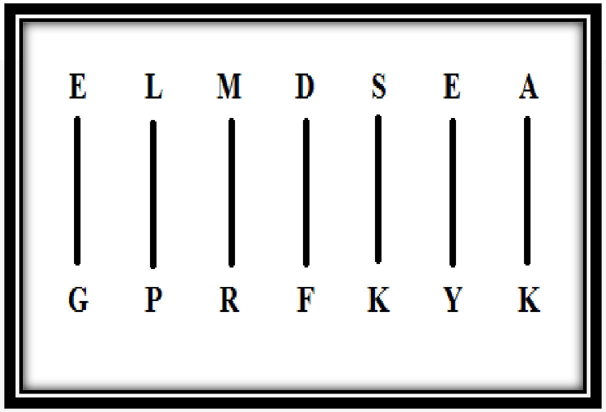
| S | i-2 | i-1 | i | i+1 | i+2 | i+3 |
|---|---|---|---|---|---|---|
| j-2 | | | | | | |
| j-1 | | s(i-1,j-1) | | | | |
| j | | | s(i,j)= s(i-1,j-1) + W(i,j) | | | |
| j+1 | | | | | | |
| j+2 | | | | | | |
| j+3 | | | | | | |
| j+3 | | | | | | |

# Protein interaction Sites Predicting Algorithm

Accepter from Psaf (-)

Donor from Cytochrome (+)

| S | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | E | L | M | D | S | E | A | A |
| | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | G | 0 | 0 | 0.2 | 0 | 0 | 0.2 | 0 | 0.2 | 0 | 0.2 |
| 2 | P | 0 | 0 | 0.2 | 0 | 0 | 0 | 0 | 0.2 | 0 | 0.2 |
| 3 | R | 0 | 0.2 | 1 | 0.4 | 0.2 | 1 | 0.2 | 1 | 0.4 | 1 |
| 4 | F | 0 | 0.2 | 0.4 | 0.78 | 0.18 | 0.4 | 0.78 | 0.4 | 0.78 | 0.6 |
| 5 | K | 0 | 0.2 | 1.2 | 0.6 | 0.98 | 1.18 | 0.6 | 1.78 | 0.6 | 1.78 |
| 6 | Y | 0 | 0 | 0.4 | 0.98 | 0.38 | 1.18 | 0.96 | 0.8 | 1.56 | 0.8 |
| 7 | K | 0 | 0.2 | 1 | 0.6 | 1.18 | 1.38 | 1.38 | 1.96 | 1 | 2.56 |
| 8 | H | 0 | 0.02 | 0.3 | 1.02 | 0.62 | 1.28 | 1.4 | 1.48 | 1.98 | 1.1 |



# Implementation

- ❑ Computational approaches were taken for predicting the interaction sites of protein pairs of cytochrome c6 and photo system I unit PsaF.

- ❑ A mathematical model is built for the interaction caused by electrostatic bond and hydrogen bond.

- ❑ Time efficient algorithms which use dynamic programming technique is designed to calculate the interaction scores and predict the interaction sites based on the scores.

- ❑ We applied the algorithm to 86 protein pairs of c6 family and PsaF family from the same organism.

- ❑ For each pair, two interaction sites with two top scores are predicted. Therefore, for 86 pairs of proteins sequence, there are totally 172 interaction sites predicted.

# Implementation

**Input**: 86 pairs of proteins from C6 and PsaF with same organisms.

**Test methods**: firstly, we predict the interaction only based on electrostatic bond then based on both hydrogen and electrostatic bonds.

**Results:** the interaction information is showed for each PsaF and c6 sequences as follows: the position of interaction site, the corresponding interaction subsequences, net charge of subsequences with the given ph value. The result of first four sequences from 86 pairs as given as follows:

# Results 1: (*For each pair of psaF and c6)*

Psaf:DIAGLTPCSESKAYAKLEKKELKTLEKRLKQYEADSAPAVALKATMERTKARFANYA KAGLLCGNDGLPHLIADPGLALKYGHAGEVFIPTFGFLYVAGYIGYVGRQYLIAVKGEAKP TDKEIIIDVPLATKLAWQGAGWPLAAVQELQRGTLLEKEENITVSPR

c6:ADLALGAQVFNGNCAACHMGGRNSVMPEKTLDKAALEQYLDGGFKVESIIYQVENG KGAMPAWADRLSEEEIQAVAEYVFKQATDAAWKY

❑ 1st interaction information:

Interaction score: 3.1

Interaction site and sequence in Psaf: 16-27 KLEKKELKTLEKR

Interaction site and sequence in c6: 65-76 DRLSEEEIQAVAE

❑ 2nd interaction information:

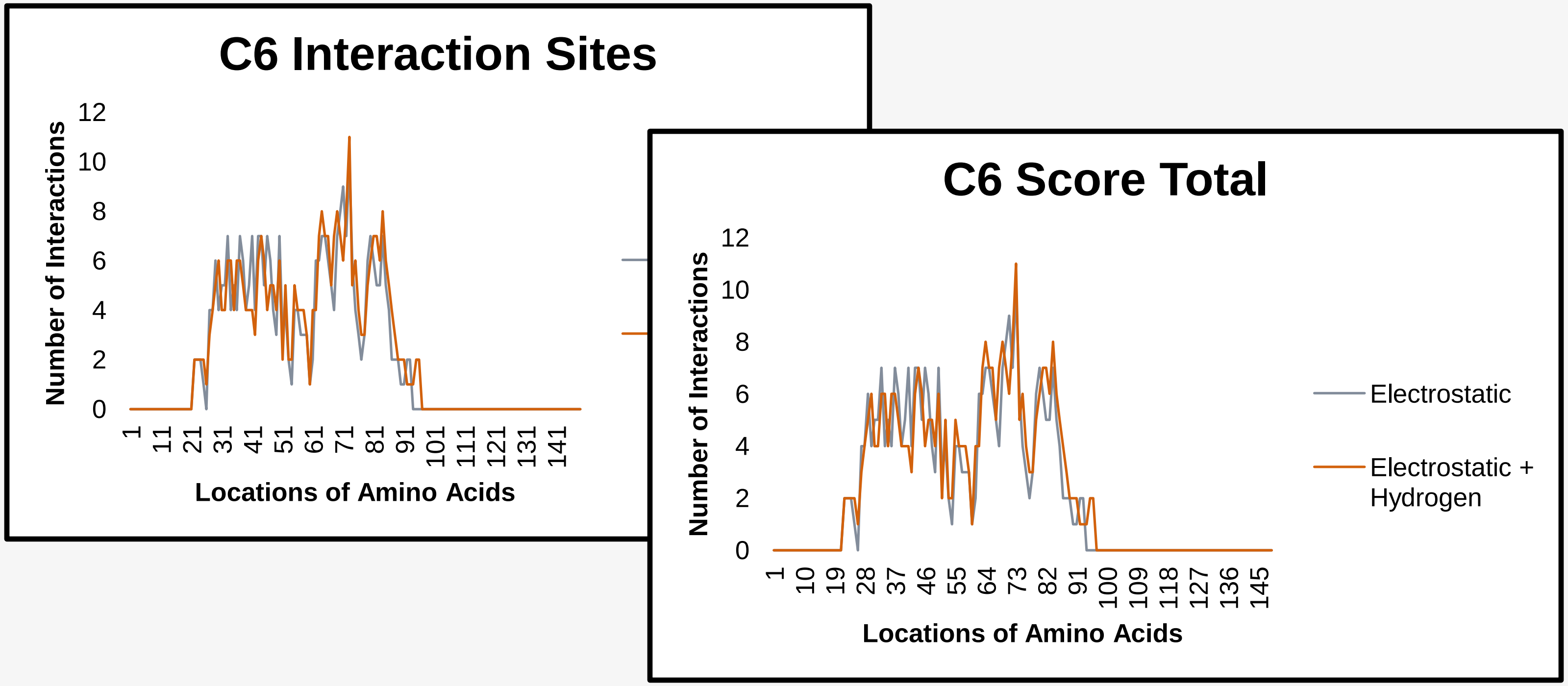
Interaction score: 2.58

Interaction site and sequence in Psaf: 19-29 KKELKTLEKRLK

Interaction site and sequence in c6: 28-38 EKTLDKAALEQY

# Result

Distribution of Interaction Sites of c6



C6 Interaction Sites

C6 Score Total

# Conclusion

- ❑ Computational approaches were proposed for predicting interaction sites of protein pairs of cytochrome c6 and photo system I unit PsaF.

- ❑ Our implementation could be used in analyzing sequences of particular interest in the evolution of protein families.

Two major tasks were achieved as follows:

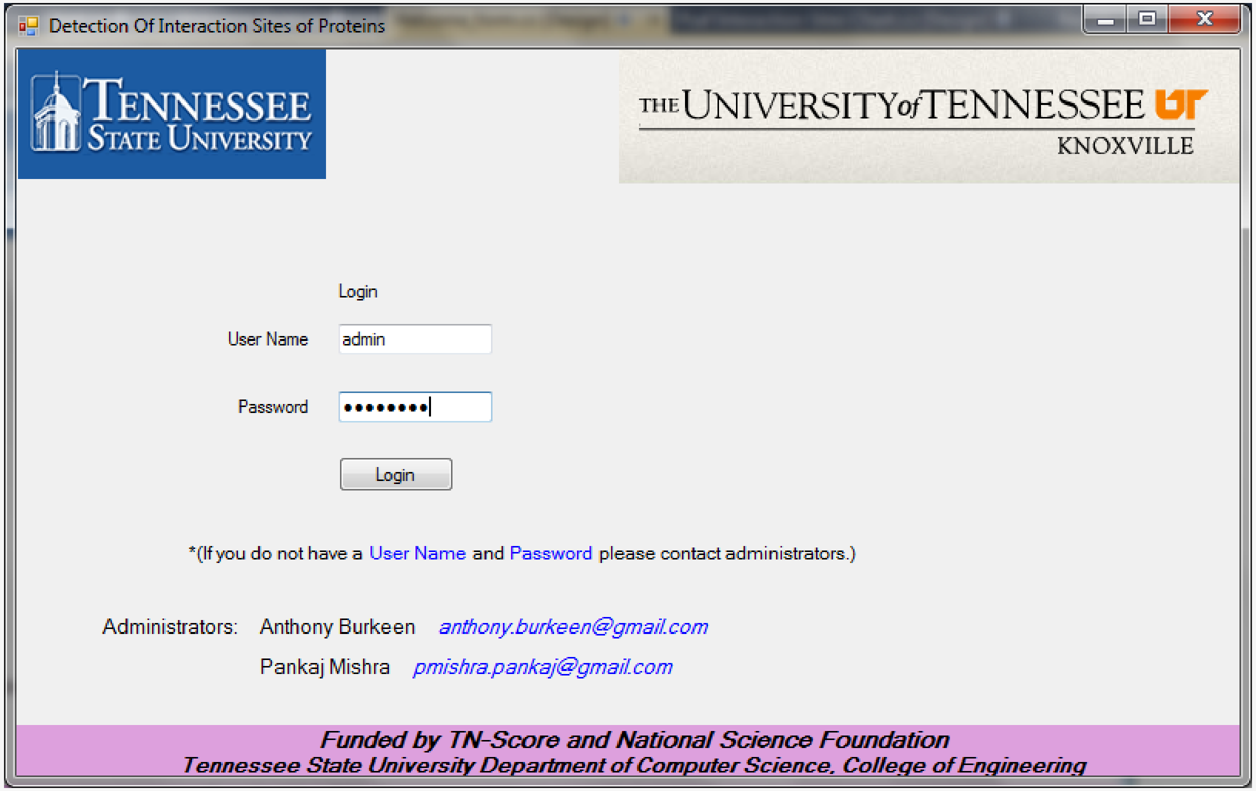
- ❑ **Task 1:** for each pair of c6 and PsaF protein sequences two interaction sites with top two score are predicted. The corresponding interaction subsequences are obtained.

- ❑ **Task 2:** the distribution of the interaction sites and score for c6 and PsaF families are investigated by statistics.

# Conclusion

- ❑In future, more issues such as hydrophobic bond, motif finding, and property in three dimensional protein structures will be considered for making the prediction more accurate, and solution will be comparison to that of laboratory experiment.

# Demonstration



Detection Of Interaction Sites of Proteins
TENNESSEE STATE UNIVERSITY
THE UNIVERSITY of TENNESSEE UT
KNOXVILLE

Login

User Name: admin

Password: ••••••••

*(If you do not have a User Name and Password please contact administrators.)

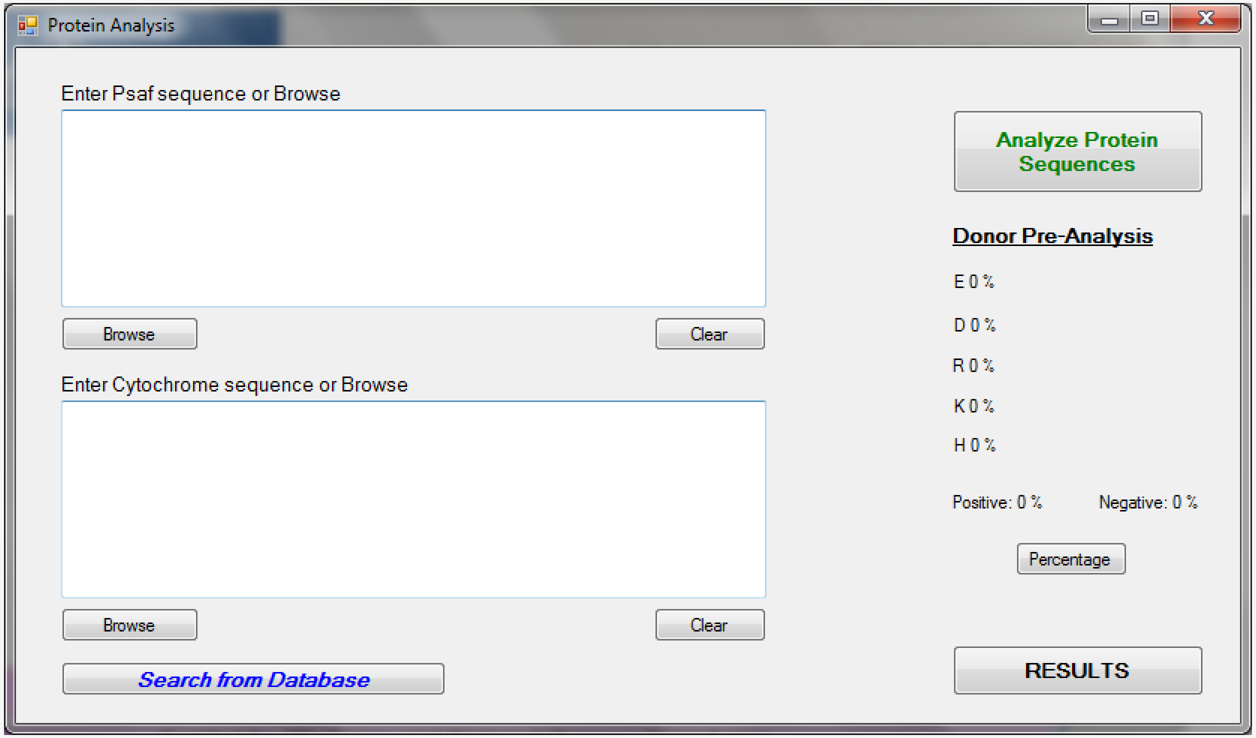
Administrators: Anthony Burkeen *anthony.burkeen@gmail.com*

Pankaj Mishra *pmishra.pankaj@gmail.com*

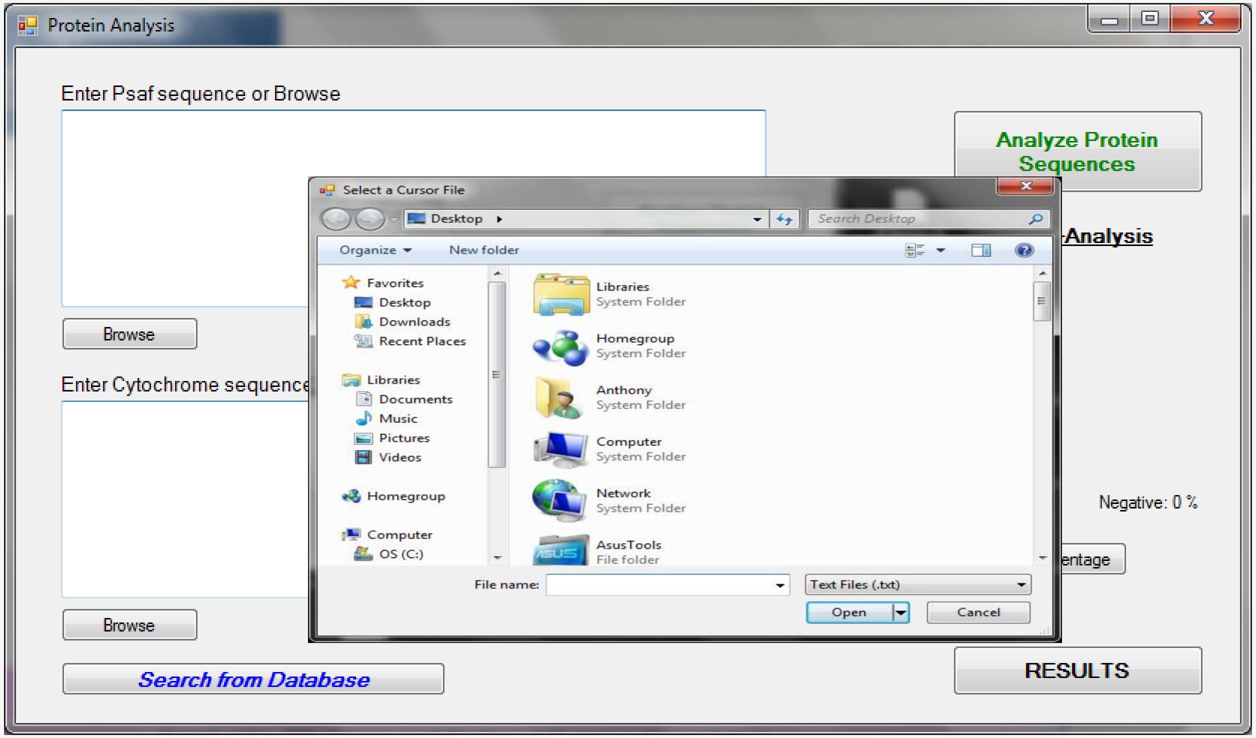
***Funded by TN-Score and National Science Foundation***

***Tennessee State University Department of Computer Science, College of Engineering***

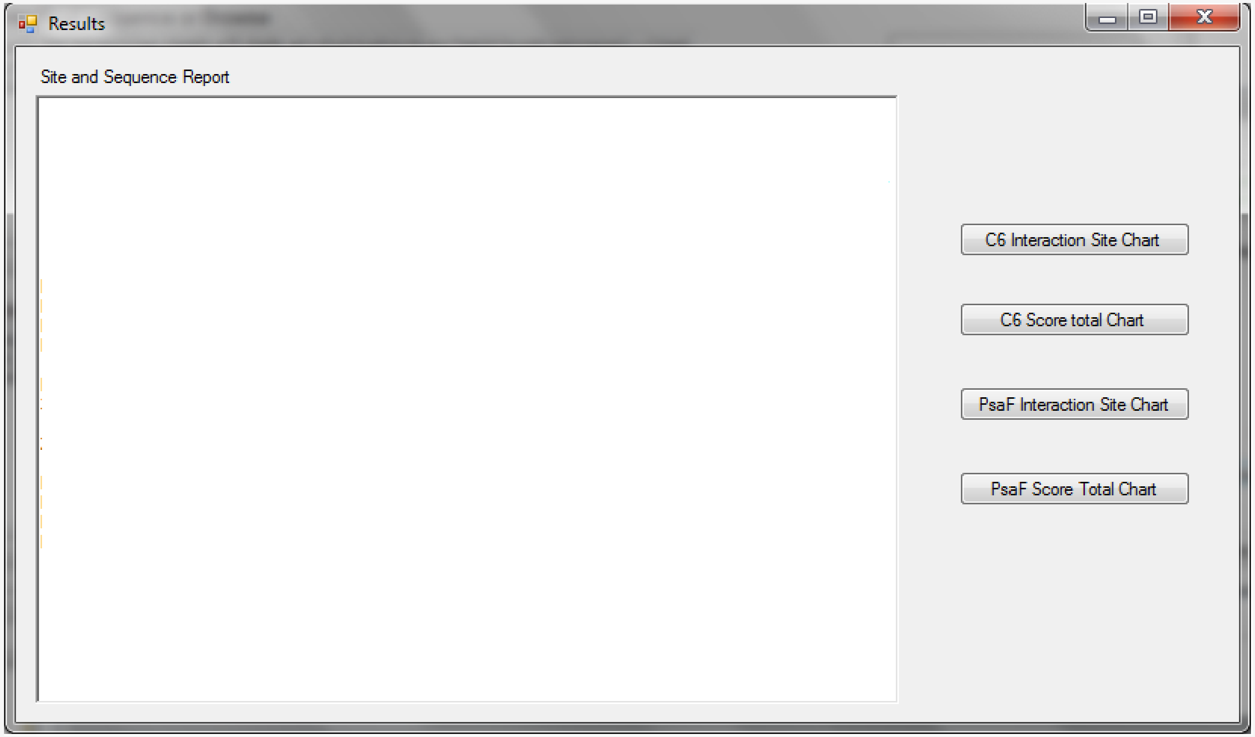
# Demonstration



# Demonstration



# Demonstration



# Acknowledgement

We are heartily thankful to our advisor Dr. Wei Chen, whose encouragement; guidance and support lead us to develop a better understanding of our project. Dr. Chen, she is always there for us to help us out in any situation.

We would also like to thank our mentor, Dr. Ali Sekmen, who always encourages us to do the best while having the good understanding of the subject and his priceless words of advice.

We will also like to thank to TN-Score and National Science Foundation for their financial support and necessary resources.

**Pankaj Mishra and Anthony Burkeen**

Joining us today

**Congratulation!! To all my classmates who are graduating.**