# Computational Approaches for Predicting Interaction Sites of Cytochrome and Photosystem I

Wei Chen, Ali Sekmen
Computer Science
Tennessee State University
Nashville, TN 37209, USA
wchen/asekmen@tnstate.edu

Barry D. Bruce, Khoa Nguyen
Biochemistry & Cellular & Molecular Biology
University of Tennessee at Knoxville
Knoxville, TN 37916, USA
bbruce/knguyen4@utk.edu

P. Mishra, L. Emujakporue, K. Wehbi
Computer Science
Tennessee State University
Nashville, TN 37209, USA
pmishra.pankaj/ejoklin/ khadijehwehbi
@gmail/yahoo/gmail.com

**Abstract**

Hydrogen is a particularly useful energy carrier for transportation. However, there are no sources of molecular hydrogen on the planet. An attractive solar-based approach is bio-hydrogen production, which utilizes protein components, Photosystem I (PSI) and cytochrome c6 (cyt c6), that function in natural photosynthesis. In aiming to increase hydrogen production, it is prudent to understand potential interactions between PSI with cyt c6, and how they affect protein-protein affinity, leading to changes in electron transfer, which would lead to overall $H_2$ yield. For this research, protein sequences from these systems are analyzed by computational approaches, in which we propose to predict the interacting residues of the cyt c6-PSI protein pair. First, the interaction relation is mathematically modeled. Then, dynamic programming algorithms are proposed to efficiently calculate the interaction score and predict the interaction sites. The proposed algorithms are applied to 86 pairs of cyt c6 and PsaF residue sequences, which have electrostatic attraction with each other. Finally, the putative interaction sites are analyzed and other chemical properties such as net charge of the residue sequences are investigated. A preliminary comparison between computational and laboratory approaches is also given.

***Keywords:*** protein-protein interaction, computational approaches, cytochrome c6, photosystem I, dynamic programming

## 1 INTRODUCTION

Solar radiation is an integral component of most renewable energy portfolios. This can be via direct photovoltaic conversion or via the production of molecular hydrogen. Hydrogen is a particularly useful energy carrier for transportation. However, there are no sources of molecular hydrogen on the planet. Thus it remains a difficult challenge to find an efficient and environmentally sustainable way of producing, capturing, storing highly attractive yet dilute energy source. The research shows that the natural process of photosynthesis can be redirected to produce molecular hydrogen [1-4]. We have characterized and partially optimized protein-metal hybrid complexes that, when exposed to light, generate hydrogen at a high rate and are temporally and thermally stable. Future improvement involves further kinetic optimization of electron transfer within photosystem I. Specifically we are using mutagenesis to increase the affinity between cyt $c_6$ and PSI from the thermophilic cyanobacterium *Thermosynechococcus elongates.* We are remodeling this protein-protein interface to include new residues that are introduced into the native complexes to create binding sites similar to those found in green algae and higher plants. The lack of a crystal structure for bound binary complex makes traditional structural biology tools unavailable to date. There have been several low resolution structural approaches such as chemical cross-linking that have investigated this interaction. For example, a mass spectrometric analysis of tryptic peptides from the cross-linked product revealed specific interaction sites between residues Lys27 of psaF and Glu69 of cyt $c_6$ and between Lys23 of psaF and Glu69/Glu70 of cyt $c_6$ [5]. Using this data, a molecular model of the intermolecular electron transfer complex is presented between eukaryotic cyt $c_6$ and PSI. This work showed that a major contributor to this interaction is electrostatic attraction between acidic residues on cyt c6 and basic residues on the luminal tail of psaF.

This paper considers the problem of optimizing the electron transfer for bio-hydrogen production by photosynthesis and outlines a purely bioinformatics approach to look at the natural variability in the occurrence of the charged residues in cyt 6 and psaF for a large set of paired proteins. Working in conjunction with this bioinformatics research, we will verify the role of these residues by mutagenesis and optimization of the binding affinity to aid the electron transfer for the two protein complexes involved (cyt $c_6$ and PSI). Protein interactions are important for all aspects of cellular functions and how they interact has become a major goal in recent years. Since laboratory methods are time and cost expensive, some computational approaches have

been proposed for predicting protein-protein interaction sites. In [6] binding hot spots in protein-protein and protein-ligand interfaces are investigated by using Q SiteFinder. It requires the protein structures for 3D grids based analysis. In [7-8] two machine learning-based methods are proposed for identifying interacting residues. The methods depend on the datasets with pre-known interaction information in learning and training processes. In [9], a probabilistic method is proposed. It requires the proteins and their motifs that can mediate the protein. Therefore, all these computational methods require either the protein structures or the datasets with pre-known interaction information. The high resolution structures of most of PSI and cytochrome proteins are unknown. Although the electrostatic and hydrophobic recognition sites on PSI are well known, the precise electrostatic recognition site on cytochrome is unknown [5].

In this research, computational approaches are proposed to identify recognition sites of binding and electron transfer in cyt $c_6$ and the PSI subunit psaF. The approaches are based on pairwise amino acid residue interaction propensities. Electrostatic bonds can directly contribute in mediating the binding affinity and further affect the electron transfer between the two proteins. First, the interaction tendencies for electrostatic and hydrogen bonds are mathematically modeled. Then, dynamic programming algorithms are designed to efficiently calculate the interaction scores and predict the interaction sites. Since the prediction is based on the interaction relations, it can be improved incrementally by adding more interaction criteria whenever necessary. The proposed algorithms are applied to 86 protein pairs from cyt $c_6$ and psaF families, which may potentially share electrostatic attractions with each other. The resulting putative interaction sites are further analyzed by other statistical methods in which chemical properties such as total net charge of residue sequences are considered. A preliminary comparison to laboratory experiments is also given.

## 2 METHODS

### 2.1 Interaction Relation Modeling

We model the interaction relation between cyt c6 and PsaF. First, electrostatic bond is considered where cyt c6 is donor and PsaF is acceptor for electron transfer. Then, hydrogen bond is considered which contributes to interaction stability. Let *A* be the set of amino acids. The amino acids with positive charges are Arginine (*R*), Histamine (*H*) and Lysine (*K*); and the ones with negative charges are Aspartic Acid (*D*) and Glutamic Acid (*E*). Let $N = \{E, D\}$ and $P = \{R, H, K\}$. An electrostatic bond based interaction relation is represented as

$R_e = \{(x, y) \mid x \in N \text{ and } y \in P\}$.

We assign weight to electrostatic bond for each pair of amino acid residues in cyt c6 and PsaF protein sequences. Given a pair of residue *x* in cyt c6 and residue *y* in PsaF, if $x \in N$ and $y \in P$, (*x*,*y*) is assigned a positive weight. *H* has 10% positive change that *R* and *K* have. Therefore, the pair contains *H* will be assigned a smaller weight. We designed two weight schemes. In one scheme, we use a window with length *l*. For pair (*x*,*y*), if $x \in N$ and $y \notin P$, (*x*,*y*) is still assigned a positive weight if *y* has a neighbor $y' \in P$ in the window, i.e., the positions of *y'* and *y* in the PsaF sequence does not exceed $\lfloor l/2 \rfloor$. It is the same for the case $x \notin N$ and $y \in P$. The weight for electrostatic bond based interaction is assigned as follows (Fig. 1):

$$W_e(x,y) = \begin{cases} \alpha & \text{if } x \in N \,\&\, y \in P' \\ 0.1\alpha & \text{if } x \in N \,\&\, y = H \\ 0.2\alpha & \text{if } x \in N \,\&\, y \in \overline{P} \,\&\, y' \in P' \\ & \text{or if } x \in \overline{N} \,\&\, y \in P' \,\&\, x' \in N \\ 0.02\alpha & \text{if } x \in N \,\&\, y \in \overline{P} \,\&\, y' = H \\ & \text{or if } x \in \overline{N} \,\&\, y = H \,\&\, x' \in N \\ -\beta & \text{Otherwise} \end{cases}$$

where *P'*={*R*,*K*}, *x'* and *y'* are the amino acid residues in window *W* of length *l*, that is, the distance of x and *x'* and y and *y'* in the window do not exceed $\lfloor l/2 \rfloor$, respectively. In Fig. 1, the window length *l* is 7.

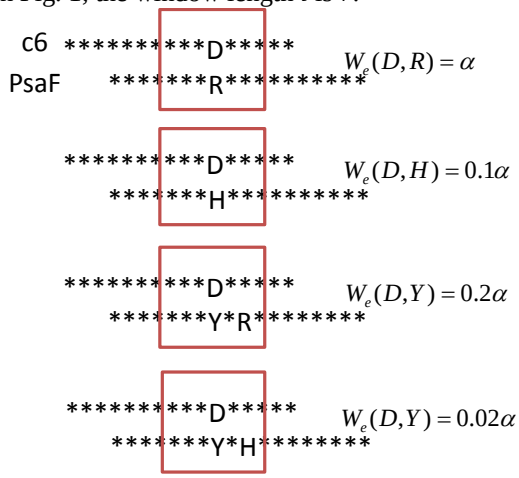


**Fig. 1** Weight scheme for electrostatic bond

Another weight scheme doesn't use window but gaps. The weight is assigned as follows:

$$W_{e'}(x,y) = \begin{cases} \alpha & \text{if } x \in N \,\&\, y \in P' \\ 0.1\beta & \text{if } x \in N \,\&\, y = H \\ -\beta & \text{Otherwise} \end{cases}$$

Arginine (R), histamine (H), lysine (K), serine (S), threonine (T), asparagine (N), glutamine (Q), tryptophan (W) and tyrosine (Y) can form hydrogen bond. A hydrogen based interaction relation is represented as

$R_h = \{(x, y) \mid x, y \in \{R,H,K,S,T,N,Q,W,Y\}\}$.

The weight for any pair in $R_h$ is assigned as follows:

$$W_h(x, y) = \begin{cases} \gamma & \text{if } (x, y) \in R_h \\ 0 & \text{otherwise} \end{cases}$$

The values of $\alpha, \beta, \gamma$ in the weight schemes can be flexible and adjusted at the test time. The total weight for any residue pair (*x*,*y*) of cyt c6 and PsaF proteins is defined as

$$W(x, y) = (W_e(x, y) \text{ or } W_{e'}(x, y)) + W_h(x, y)$$

Net charge is another chemical property which may relate to electron transfer. Since we consider the attraction between two putative interaction residue sequences, we calculate the net charge based on side chains but not C and N terminals. The amino acids having positive *pKa* values are *R, H, K*, which are 12.48, 6.10 and 10.53, respectively. The amino acids having negative pKa values are *D, C, E*, and *Y*, which are -3.86, -8.00, -4.07 and -10.07, respectively [10]. The value of ph changes from 6.25 to 8 with interval 0.25.

### 2.2 Algorithms

Given a pair of cyt c6 and PsaF protein sequences $X = x_1 x_2 \cdots x_m$ and $Y = y_1 y_2 \cdots y_n$, the algorithms in this section use dynamic programming technique to predict the interaction sites and extract the corresponding interaction residue subsequences *X'* from *X* and *Y'* from *Y*. The algorithms use an $n \times m$ matrix to calculate the score of interaction for any pair of subsequences in *X* and *Y*; then track back from the location with the highest score to get the pair of interaction residue subsequences.

We designed two algorithms. One uses the weight scheme with window. In the algorithm, the score at $(x_i, y_j)$ is decided by the score at $(x_{i-1}, y_{j-1})$ and weight $W(x_i, y_j)$, where $W(x, y) = W_e(x, y)$ if only uses electrostatic bond and $W(x, y) = W_e(x, y) + W_h(x, y)$ if it uses both electrostatic and hydrogen bonds. The score is calculated as follows:

$$S[i, j] = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0 \\ \max\{S[i\text{-}1,j\text{-}1] + W(x_i, y_j),\ 0\}, & \\ & \text{Otherwise} \end{cases}$$

The algorithm finds top *k* scores, and then track back to get all *k* interaction sites and corresponding pairs of interaction subsequences. Fig. 2 shows an example of calculating matrix *S*, where X = AELMDSEAE and Y = GPRFKYKH, $W(x, y) = W_e(x, y)$, window length =7, $\alpha$ = 1, and $\beta$ = 0.22. Since the window length is 7, when calculating *S*[4,5], the window of ELMDSEA and GPRFKYK is used. The largest score is 2.56. By tracking back from that location until the location with score 0 we get the corresponding interaction subsequences:

DSEAE
RFKYK.

The second largest score is 1.98. By tracking back from that location to the location with score 0, we get the interaction subsequences

ELMDSEA
PRFKYKH.

The algorithm is given as follows:

**Algorithm 1 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 2: Find *k* top score in *S*
Step 3: Track back to find *k* interaction subsequences

Another algorithm uses the weight scheme without window and allows the interaction subsequences have gaps. The score at $(x_i, y_j)$ is decided by the score at $(x_{i-1}, y_{j-1})$, $(x_{i-1}, y_j)$, $(x_i, y_{j-1})$ and weight $W(x_i, y_j)$, where $W(x, y) = W_{e'}(x, y) + W_h(x, y)$ or $W_{e'}(x, y)$. The score is calculated as follows:

$$S[i, j] = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0 \\ \max\{S[i\text{-}1,j\text{-}1] + W(x_i, y_j),\ S[i\text{-}1,j] - g, & \\ \quad S[i,j\text{-}1] - g,\ 0\}, & \text{Otherwise} \end{cases}$$

**S**

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | A | E | L | M | D | S | E | A | E |
| | 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 | 0.2 | 0 | 0.2 | 0 | 0.2 |
| 3 | R | 0 | 0.2 | 1 | 0.4 | 0.2 | 1 | 0.4 | 1 | 0.4 | 1 |
| 4 | F | 0 | 0 | 0.4 | 0.78 | 0.18 | 0.4 | 0.78 | 0.6 | 0.78 | 0.6 |
| 5 | K | 0 | 0.2 | 1 | 0.6 | 0.98 | 1.18 | 0.6 | 1.78 | 0.8 | 1.78 |
| 6 | Y | 0 | 0 | 0.4 | 0.78 | 0.38 | 1.18 | 0.96 | 0.8 | 1.56 | 1 |
| 7 | K | 0 | 0.2 | 1 | 0.6 | 0.98 | 1.38 | 1.38 | 1.96 | 1 | 2.56 |
| 8 | H | 0 | 0.02 | 0.3 | 1.02 | 0.62 | 1.08 | 1.4 | 1.48 | 1.98 | 1.1 |

**Fig. 2** Interaction score matrix *S*

Fig. 3 shows an example of calculating score without using window, where $W(x, y) = W_{e'}(x, y)$, $\alpha$ = 1, $\beta$ = 0.22, and $g$ = 0.2. The value of $g$ is the penalty for a gap. In the matrix *S*, The largest score is 2.56. The arrows show where the value of *S*[*i*,*j*] comes from. By tracking back from that location with the largest score until the location with score 0 we get the corresponding interaction subsequences:

DSEAE
RFKYK.

The second largest score is 2.36. By tracking back from that location to the location with score 0, we get the interaction subsequences:

ELMDSE
R– FKYK.
where there is a gap between R and F.

S

| | | | 1 A | 2 E | 3 L | 4 M | 5 D | 6 S | 7 E | 8 A | 9 E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | P | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | R | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 4 | F | 0 | 0 | 0.8 | 0.78 | 0.58 | 0.8 | 0.78 | 0.8 | 0.78 | 0.8 |
| 5 | K | 0 | 0 | 1 | 0.80 | 0.60 | 1.58 | 1.38 | 1.78 | 1.58 | 1.78 |
| 6 | Y | 0 | 0 | 0.8 | 0.78 | 0.58 | 1.38 | 1.36 | 1.58 | 1.56 | 1.58 |
| 7 | K | 0 | 0 | 1 | 0.78 | 0.58 | 1.58 | 1.38 | 2.36 | 2.16 | 2.56 |
| 8 | H | 0 | 0 | 0.8 | 0.78 | 0.58 | 1.38 | 1.36 | 2.16 | 2.14 | 2.36 |

**Fig. 3** Interaction score matrix without using window

The algorithm is as follows:

**Algorithm 2 PredicUsingGap (*X,Y,W*)**

Same as Algorithm 1 except

$S[i,j] = \max\{S[i,j]+W(i,j), S[i-1,j]-g, S[i,j-1]-g, 0\}$

In Algorithm 1 and Algorithm 2, $k$ interaction subsequences are picked without overlap. The time complexity of the algorithms are O($mn$).

**Remark:** Algorithm1 and Algorithm 2 are time optimal. It can be speed up by using multiple processors. Theoretically, the algorithms can be similarly executed in O(log $m$ log $n$) time using O($mn$ / log $m$) processors in CREW PRAM model by A. Apostolico *et al.*'s approach, where $m = \min\{|X|, |Y|\}$, $n = \max\{|X|, |Y|\}$ and $X$ and $Y$ are the pair of protein sequences [11], and O(1) time using $m + n$ processors in BSR model [12]. Practically, we can using a computer with multiple cores to speed up the algorithms as follows: divide the $|X| \times |Y|$ matrix $S$ in to $k \times k$ blocks such that each block have $|X|/k \times |Y|/k$ elements and can be calculated in O($|X|/k \times |Y|/k$ ) time by one processor. First, calculate the blocks in the first diagonal are calculated, then the ones in the second diagonal, until the ones in the (2$k$–1)th diagonal. Since the calculation on the $i$th diagonal only depends on the values in the ($i$–1)th diagonal. Each block on the same diagonal can be calculated in parallel. Therefore, the problem can be solved in $O((2k-1)(mn/k^2)) = O(mn/k)$ time with $k$ processors, where $1 \le k \le m$.

S

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | ... | k |
| 2 | 3 | ... | k | K+1 |
| 3 | ... | k | K+1 | ... |
| ... | k | K+1 | ... | 2k-2 |
| k | K+1 | ... | 2k-2 | 2k-1 |

**Fig. 4** $k$-division of matrix S

## 3. RESULTS

### 3.1 Input and Output

The input dataset, parameters and the output are set as follows:

1) Totally, 86 pairs of protein sequences from cyt c6 and PsaF are used for the test. The datasets are given from Dr. Bruce's Lab an each pair belongs to the same organism and is able to have electrostatic attractions with each other.
2) For each pair of sequences, three interaction sites which have top three scores and corresponding pairs of interaction subsequences are predicted.
3) In weight schemes $W_e$ and $W_{e'}, \alpha = 1, \beta = 0.22$. In weight scheme $W_h, \gamma = 0.1$. In Algorithm 2, $g$ = 0.2.
4) For each pair of protein sequences, the original sequences, three interaction sites with the scores, corresponding interaction subsequences, and net charge of each subsequence are output as follows:
   Psaf:MRRLFALILAIGLWFNFAPQAQALGANLVPCKD
   SPAFQALAENARNTTADPESGKKRFDRYSQALCGPE
   GYPHLIVDGRLDRAGDFLIPSILFLYIAGWIGWVGRA
   YLQAIKKESDTEQKEIQIDLGLALPIISTGFAWPAAAI
   KELLSGELTAKDSEIPISPR
   c6:MENVGCEENLLRLILVNLLLVIALLCNLTIIYPALA
   AETSNGSKIFNANCAACHIGGANILVEHKTLQKSGLS
   KYLENYEIEPIQAIINQIQNGKSAMPAFKNKLSEQEIL
   EVTAYIFQKAETGW
   1st interaction site information:
   Interaction score: 2.76
   Interaction site location and subsequence in Psaf:
   54-59, u = KKRFDR
   Interaction site and subsequence in c6:
   106-111, v = EQEILE
   Net charge:
   when ph = 6.25 net charge for u = 3.00395114057246,
   net charge for v = -2.98030929177886
   when ph = 6.5 net charge for u = 3.00209690387591
   net charge for v = -2.98889520136613
   ……..

The datasets and output can be found at webpage www.tnstate.edu/faculty/wchen/research.aspx by clicking the title of this paper in publication section.

### 3.2 Comparison of the Algorithms

We compare Algorithm 1 PredicUsingWindow and Algorithm 2 PredicUsingGap. For the simplicity, we use the electrostatic bond only in the weight schemes. From the results, we found that Algorithm 2 tends to give the interaction sites that have the same number of the positive charged and negative charged residues. For example, for the pair of protein sequences in Section 3.1, the first interaction site and the corresponding interaction residue subsequences predicted from Algorithm 1 are

PsaF: 54-59, u = KKRFDR
cyt c6: 106-111, v = EQE ILE

and the results from Algorithm 2 are

PsaF: 55-59, K_RFDR

cyt c6: 106-111, EQEI LE.

In the first pair of subsequences, there are five positive charged residues (KKRR) and four negative charged residues (EEE), and in the second pair of subsequences, there are three positive (KDR) and three negative (EEE) charged residues. Therefore, the algorithm can be selected based on the property to be investigated.

### 3.3 Comparison of Laboratory and Computational Approaches

Very little results are known from Laboratory work for electrostatic recognition sites between cyt c6 and the photosystem I subunit PsaF. In [6], Mass Spectrometry is used to precisely identify electrostatic recognition sites between the following protein pairs of cyt c6 and PsaF sequences (The first more than twenty amino acid residues have been skipped from the original protein sequences) :

Psaf:DIAGLTPCSESKAYAKLEKKELKTLEKRLKQYEADS APAVALKATMERTKARFANYAKAGLLCGNDGLPHLIAD PGLALKYGHAGEVFIPTFGFLYVAGYIGYVGRQYLIAVK GEAKPTDKEIIIDVPLATKLAWQGAGWPLAAVQELQRGT LLEKEENITVSPR

c6:ADLALGAQVFNGNCAACHMGGRNSVMPEKTLDKAA LEQYLDGGFKVESIIYQVENGKGAMPAWADRLSEEEIQA VAEYVFKQATDAAWKY.

The laboratory approach shows that the cross-lined interaction happens in following interaction subsequences:

PsaF: 21-28, ELKTLEKR

cyt c6: 67-81, LSEEEIQAVAEYVFK

We use the computation approaches to predict the interaction site of the same pair. The result from Algorithm 1 is

PsaF: 22-29, KTLEKRLK

cyt c6: 64-70, DRLSEE _E

and result from Algorithm 2 is

PsaF: 15-27, KLEKKELKTLEKR

cyt c6: 64-76, DRLSEEEIQAVAE.

Both algorithms accurately predict the interaction site and the corresponding subsequences.

### 3.4 Distribution of Interaction in PsaF and Cyt c6

One of the purposes in this research is to investigate the locations that electron transfer most possibly happens in PsaF and cyt c6 protein families. The algorithms predict three top interaction sites for each pair of PsaF and c6 protein sequences. Therefore, there are 258 interaction subsequences predicted from 86 sequences of PsaF and cyt c6, respectively. In order to show the distribution, for each position in the sequences of cyt c6 (PsaF), we count the number of times it involves and total score it receives in 258 putative interaction subsequences. The number of interactions and total score at each position $i$ are defined as follows: numInt($i$) = |S($i$)|, and tolScore($i$) = $\sum_{\forall s \in S(i)}$ score of $s$ , where $S(i)$ is the subset of 258 interaction subsequences in PsaF (cyt c6) and $s \in S(i)$ if position $i$ is included in $s$. Due to the space limitation, we only show the result from Algorithm 1. To avoid the noise we skip the first five positions for all sequecnes. In Fig. 5, the blue line shows the distribution of interaction numbers at each position in PsaF sequences when using electrostatic bond in the weight scheme and the red line shows the distribution when using both electrostatic and hydrogen bond in the weight scheme. We can see that in both cases the positions surround 43, 57, 78 and 105 reach the peaks. Fig. 6 shows that distribution of interaction score in PsaF. In the blue line the peaks are similar to these in Fig. 5. In the red line the peak at 78 shits to 97.

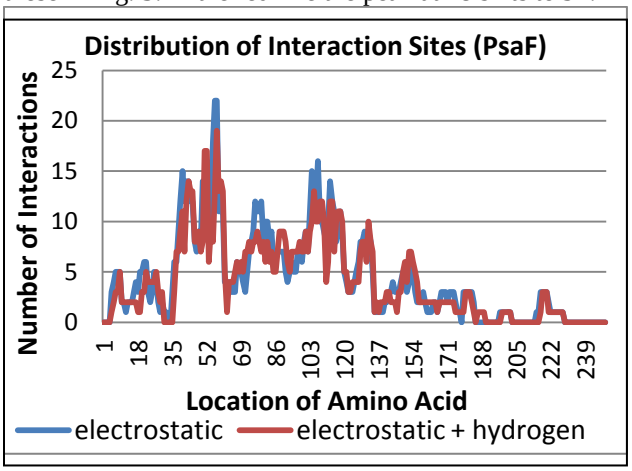


**Fig. 5** Distribution of interaction Sites in PsaF

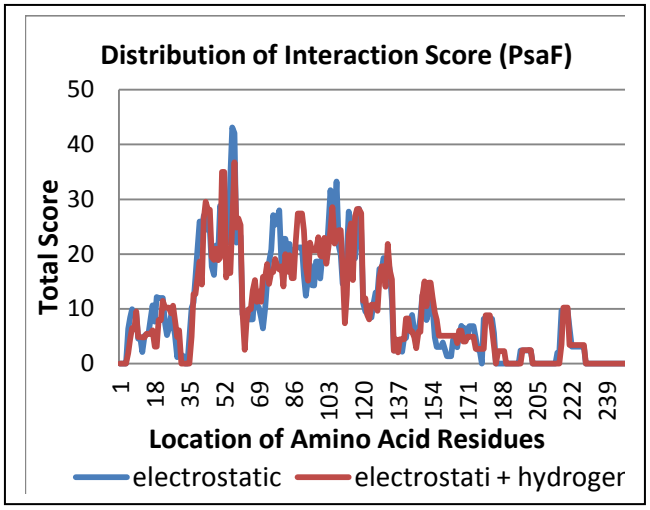


**Fig. 6** Distribution of interaction score in PsaF

Fig. 7 shows the distribution of both interaction number and score for cyt c6. It shows no matter using electrostatic bond or using both electrostatic and hydrogen bonds the number and score of interactions at the positions surround 71, 109, and 148 reach the peaks.

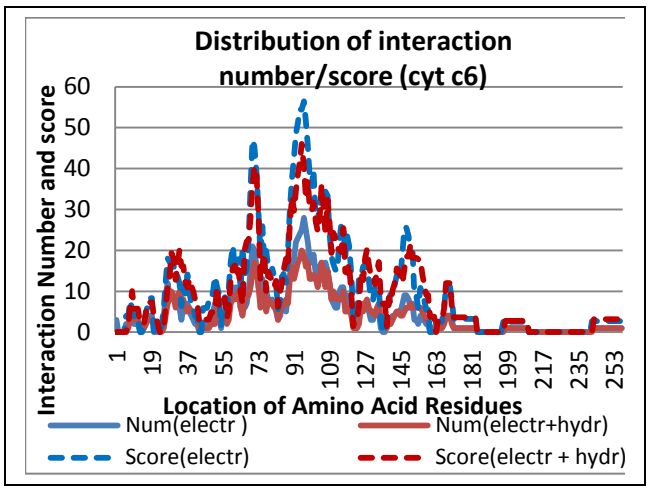


**Fig. 7** Distribution of interaction number and score in cyt c6

### 3.5 Net Charges and Inference to Interaction

The relation of net change and interaction is investigated. At each position *i* in Psa F (cyt c6) sequence *s*, the net charge *s*(*i*) is calculated from *ph* = 6.25 to *ph* = 8 at each interval 0.25 use a window of length 7 as follows: s(*i*) = the net charge of subsequence of *s* from position *i* – 3 to *i* + 3, and the net charge at position *i* is defined as $CH(i) = \sum_{\forall s \in S} s(i)$, where *S* is the given 86 PsaF (cyt c6) sequences. Fig. 8 shows the net charges for *ph* = 6.25 and *ph* = 8. We can see that the positions surround 43, 57, 78, and 105 in PsaF reaching the peaks in Fig. 5 also reaching

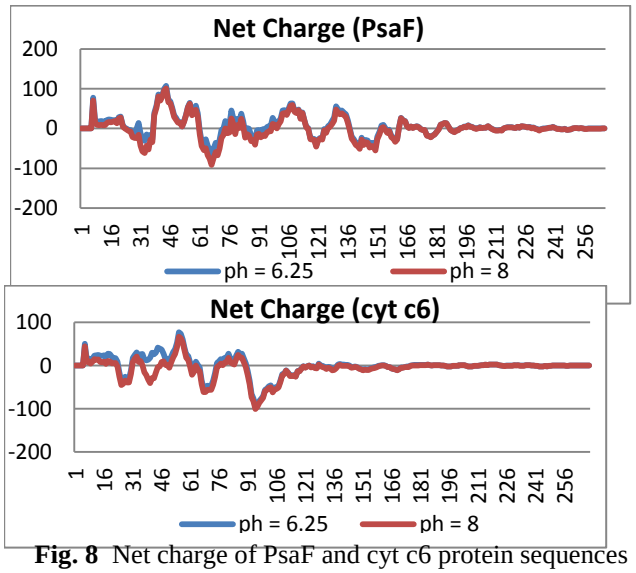


**Fig. 8** Net charge of PsaF and cyt c6 protein sequences

the positive peaks in Fig. 8. The positions surround 71 and 109 in cyt c6 that reach the peaks in Fig. 7also reach the negative peaks in Fig. 8. However, position 148 doesn't reach the negative peak.

## 4. CONCLUSION

We proposed the mathematical model and computational approaches for predicting interaction sites of Cytochrome and Photosystem I. In the future, we will add more interaction criteria into the model and algorithms. We will also find more laboratory results to compare with the results from computational approaches.

**Acknowledgement:** The research is partially supported by the grants NSF 1137484 and TN-SCORE, 2012-2013.

## REFERENCES

1. IJ., Iwuchukwu, M. Vaughn, N. Myers, H. O'Neill, P. Frymier, B. Bruce, "Self-organized photosynthetic nanoparticle for cell-free hydrogen production," Nature Nanotechnology 5, 1, 73-19, January, 2010.
2. K. Matsumoto, V. Vaughn, B.D. Bruce, S. Koutsopoulos, S. Zhang, "Designer Peptide Surfactants Stabilize Functional Photosystem-I Membrane Complex In Aqueous Solution For Extended Time", J. Phys. Chem B. 113:1, 75-83, 2009.
3. I.J. Iwuchukwu, M. Myers, M. Vaughn, H. O'Neill, P. D. Frymier, B. Bruce, "Thermotolerant photodependent hydrogen production by platinized photosystem I reaction centers and recombinant cytochrome c553 from thermosynechococcus elongatus", Annual Fall Meeting of the AIChE, Philadelphia, PA, 2008.
4. "Optimizing electron transfer for biohydrogen Production by photosynthesis," http://web.utk.edu/~seerc/mainproj1.html#fy2009.1.
5. F. Sommer, F. Drepper, W. Haehnel, M. Hippler, "Identification of precise electrostatic recognition sites between cytochrome c6 and the photosystem I subunit PsaF using mass spectrometry," Journal of Biological Chemistry Vol. 281, No. 46, pp. 35097-35101, 2006.
6. N. J. Burgoyne, R. M. Jackson, "Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces," Bioinformatics, Vol. 22, no. 11, pp. 1335 – 1342, 2006.
7. Y Ofran, B. Rost, "ISIS: Interaction sites identified from sequence," Bioinformatics," Vol. ECCB, pp. e13 – e16, 2006.
8. H. Wang, E. Segal, A. Ben-Hur, Q-R Li, M. Vidal, D. Koller, "InSite: A computational method for identifying protein-protein interaction binding sites on a proteome-wide scale," Benome Biology, 8:R192, 2007.
9. M. Sikic, S. Tomic, K. Vlahovicek, "Prediction of protein-protein interaction sites in sequences and 3D structures," PLoS Computational Biology, Vol. 5, Issue 1, pp. 1-9, 2009.
10. D. S Moore, "Amino acid and peptide net charges: a simple calculation procedure," Biochemical Education, 13(1), pp. 10 – 11, 1985.
11. A. Apostolico, M. J. Atallan, L. Larmore, S. McFadding, "Efficient parallel algorithms for string editing and related problems," SIAM J. Comput., Vol. 19, Issue 5, pp. 968-988, 1990.
12. J-F. Myoupo, D. Seme, "Time-efficient parallel algorithms for the longest common subsequence and related problems," Journal of Parallel and Distributed Computing, Vol. 57, pp. 212-223, 1999.