

# Novel Motif Detection Algorithms for Finding Protein-Protein Interaction Sites

January Wisniewski

MS in Computer Information System Engineering

Advisor: Dr. Chen

College of Engineering, Department of Computer Science

Tennessee State University

Spring 2014

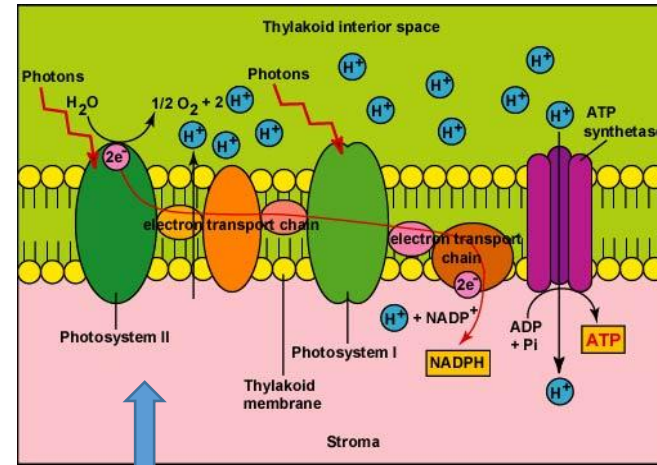
This work is supported by a collaborative contract from NSF and TN-SCORE

# Outline

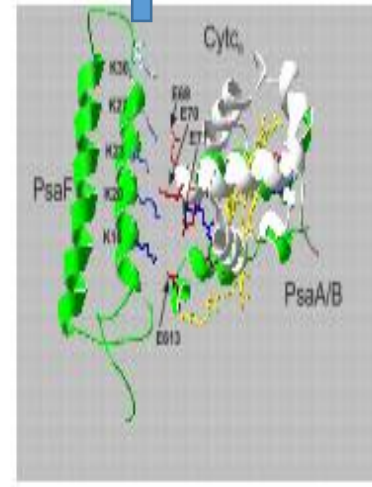
- Research Background
- Problem Statement
- Challenges
- Approach
- Incremental Design of Algorithms
- Testing and Evaluation
- Summary and Future work

# Research Background - Motivation

- Hydrogen is 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)
- 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<sub>2</sub> yield.



Natural photosynthetic process is not efficient and quantitative !!!



Artificial photosynthetic process: by adding the proteins that can donate and accept large number of electrons, can increase the production of hydrogen.

# Research Background – Why Computational Approach?

## ➤ Biologist's Approach

- Due to the lack of a crystal structure for bound binary complexes, traditional structural biology tools are rendered unavailable to date.
- Even when the Biologist's approaches are developed, they are expensive and time consuming.

## ➤ Computer Scientist's Approach

- Predict the candidates for the Biologist
- Resource and time efficient

# Research Background – What We Have Done

**Previous work:** Computational approaches have been 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, hydrogen bonds and hydrophobic bonds** are mathematically modeled and used for interaction prediction algorithms

## **Question:**

In genetics, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance or functionality.

Will the motifs also play a role in protein-protein interaction?

# Problem Statement

- This research addresses the problem of computationally predicting the interaction sites of protein pairs (donors and acceptors) that tap into photosynthetic processes to produce efficient and inexpensive hydrogen
- More specifically, we are attempting to use motifs to make more accurate predictions of the interaction sites between Cyt  $c_6$  and the PSI subunit PsaF.

# Challenges

- Motif detection requires an exhaustive search method, making it an NP-hard problem. Meaning that it is unrealistic to find the optimal solution when the problem size is large.
- For this research, we need to detect the motifs from 86 amino acid sequences from both PsaF and Cyt c6. Meaning that the size of the problem is large.

# Approach – Incremental Design

Incrementally improving algorithms to increase the score of motif candidates

## Score of a candidate of motif

GGGCT	<u>A T C C A G C T</u>	GGGTCGTCACATTCCCCTTTTCGA
TGAGGGTGCCCAATAA	<u>G G G C A A C T</u>	CCAAAGCGGACA
	<u>A T G G A T C T</u>	GATGCCGTTTGACGACCTAAATCAACGG
GG	<u>A A G C A A C C</u>	CCAGGAGCGCCTTTGCTGGTTCTACC
TTTTCTAAAAAGATTATAATGTCGGTCC	<u>T T G G A A C T</u>	
GCTGTACACACTGGATCATGCTGC	<u>A T G C C A T T</u>	TTCA
CATGATCTTTTG	<u>A T G G C A C T</u>	TGGATGAGGGAATGAT

A	5	1	0	0	5	5	0	0
T	1	5	0	0	0	1	1	6
G	1	1	6	3	0	1	0	0
C	0	0	1	4	2	0	6	1

**Positions of motif = (6,17,1,3,29,25,13)**  
**Score(s, DNA) = 5+5+6+4+5+5+6+6 = 42**

**Consensus**      **A T G C A A C T**



# Incremental Design of Algorithms: Brute Force

## Brute Force for Motif Finding Problem

Let  $p$  be a set of  $l$ -mers from  $t$  DNA sequences and the  $l$ -mers start at the position  $s = (s_1, s_2, \dots, s_t)$ . Find  $p$  which has the maximum  $Score(s, DNA)$  by checking all possible position  $s$ .

## BruteForce-MotifFinding(DNA, t, n, l)

```
bestScore := 0;
for  $i1 := 1$  to  $n-l+1$ 
  for  $i2 := 1$  to  $n-l+1$ 
    .....
    for  $it := 1$  to  $n-l+1$ 
       $S = (i1, i2, \dots, it)$ 
      if ( $Score(S, DNA) > bestScore$ )
         $bestScore := Score(S, DNA)$ 
         $bestMotifPosition = S$ 
return  $bestScore$  &  $bestMotifPosition$ ;
```

**DNA:** DNA sequences  
**t:** number of DNA sequences  
**n:** length of DNA sequences  
**l:** length of the motif

Time Complexity :  $O(n^t lt)$

# Incremental Design of Algorithms: Greedy/Heuristic

## Greedy-MotifFinding(DNA, t, n, l)

bestMotif := (1,1,...,1);

s := (1,1,...,1)

for s1 := 1 to n-l+1

for s2 := 1 to n-l+1

S := (s1, s2, 1, ..., 1)

if (Score(S, Seq) > bestScore)

bestScore := Score(S, DNA);

bestMotif Position := S

for i := 3 to t

for si := 1 to n-l+1

S := (s1, s2, ..., si, 1, ..., 1)

if (Score(S, DNA) > bestScore)

bestScore := Score(S, Seq);

bestMotif Pos := S;

return bestScore & bestMotifPos

## Greedy Algorithm for Motif Finding Problem

**Step 1 (initialization)** Assume that all motifs in the sequence start from the first position.

**Step 2** Find the *l*-mers locally optimal in the first two sequences (the motifs in other sequences are fixed).

**Step 3** For  $i = 1$  to  $t$ , find the *l*-mer locally optimal in  $i$ th sequence when the motifs in other sequences are fixed.

Time Complexity:  $O(n^2tl + nt^2l)$

**Weakness: It can fall into local optimality**

# Incremental Design of Algorithms: Improved Heuristic

## ImprovedGreedy-MotifFinding(DNA, t, n, l)

```
lastBestScore := 0; bestScore := 1;
while (bestScore > lastBestScore)
{
    Greedy-MotifFinding(DNA, t, n, l)
    {
        ....
    }
    return bestScore and bestMotifPos;
}
```

### Improved Greedy for motif finding

Repeat executing Heuristic Algorithm until the score of  $l$ -mers cannot be improved.

Time Complexity:  $O(k(n^2tl + nt^2l))$ , where  $k$  is the repeat times.

# Incremental Design of Algorithms: Divide and Conquer

```
DivideConquer(DNA[i..j], t, n, l)
if (j-i) < 4
  return Greedy(DNA[i..j], t, n, l)
else
  k = (i+j-1)/2
  x = DivideConquer(DNA[i..k], t, n, l)
  y = DivideConquer(DNA[k+1..j], t, n, l)
  if x.score > y.score
    improve DNA[k+1..j] by the motifs in DNA[i..k]
    with greedy/heuristic technique
  else
    improve DNA[i..j] by the motifs in DNA[k+1..j]
    with greedy/heuristic technique
  return bestScore and bestMotifPosition
```

## Divide-and-Conquer for Motif Finding Problem

### Divide Step

Divide the set of sequences into half and half.

### Conquer Step

- (1) Recursively find the *l*-mers locally optimal in the first half of sequences.
- (2) Recursively find the *l*-mers locally optimal in the second half of sequences.

### Merge Step

If the score of the motif from the first half is larger than that from the second half, use the first to improve the second one; otherwise use the second one to improve the first one.

Time Complexity :

$$T(n) = 2T(n/2) + nt^2l/2 \quad \text{if } t > 4$$

$$= n^2tl \quad (\text{use greedy}) \quad \text{if } t \leq 4$$

$$T(n) = O(n^3tl)$$

# Testing and Evaluation: Sample Data

Input: 7 DNA sequences of length 36

Output: the candidate of motif with length 8

Algorithms	Score of Motif	Position of Motif	Running Time
Brute Force			Years
Greedy	68	10, 27, 0, 11, 8, 8, 10, 26, 0, 2, 0, 2, 1, 2	3.46 ms
Improved Greedy	72	10, 26, 0, 2, 8, 8, 10, 26, 1, 2, 0, 2, 1, 2	5.19ms
Divide-and-Conquer	86	25, 2, 10, 23, 23, 23, 25, 2, 25, 6, 10, 15, 25, 6	2.006 s

# Testing and Evaluation: Experiment Data

Input: 86 PSI PsaF protein sequences & 86 Cyt c6 protein sequences

Output: Motif candidates of PsaF sequences & c6 sequences

## Sample of PsaF protein sequences:

1.ANLVPCKDSPAFAQALAEARNTTADPESGKKRFDRYSQLCGPEGYPHLIVDGRLLDRAGDFLIPSILFLYIAGWIGWVGRAYLQAIKKESDTEQKEIQIDLGLALPIISTGFAWPAAAIKELLSGELTAKDSEIPIPR

2.DIGGLVPCSESPKFOERAAKARNTTADPNSGQKRFEYSSALCGPEDGLPRIIAGGPMRRAGDFLIPGLFFIYIAGGIGNSSRNYQIANRKKNAKNPAMGEIIVPLAVSSTIAGMAWPLTAFRELTSGELTVPDSDTVSPR

3.LCGPEDGLPRIIAGGPWSRAGDFLIPGLLFIYIAGGIGNASRNYQIANRKKKNPKNPAMGEIIVPLALSSTIAALAWPVKALGEVTSGLKTVPDSDTVSPR

4.ADLTPCAENPAFQALAKNARNTTADPQSGQKRFEYSSALCGPEGYPHLIVDGRLLDRAGDFLIPSILFLYIAGWIGWVGRAYLQAIKKDSDETEQKEIQIDLGLALPIIATGFAWPAAAVKELLSGELTAKDSEITVSPR

5.DISGLTPCKDSKQFAKREKQQIKKLESSLKYAPESAPALALNAQIEKTKRRFDNYGKYGLLCGSDGLPHLIVNGDQRHWGEFITPGILFLYIAGWIGWVGRSYLIAISGEKKPAMKEIIVPLASRIIFRGIWPVAAYREFLNGDLIAKD

.....

## Results:

**Efficiency:** The candidates of the motif of 86 PsaF protein sequences and the motif of 86 c6 protein sequences were efficiently calculated by the proposed algorithms.

**Effectiveness:** There are 23 different amino acids in a protein sequence instead of 4 different nucleotide bases; therefore, the score as determined by the appearance of amino acids is not as reliable because of the lower average frequency of its components.

# Summary and Future Work

## Summary

- ✓ Designed a number of algorithms which incrementally improved the score of candidates of motifs.
- ✓ Implemented, tested, and evaluated the algorithms using 86 PSI PsaF and Cyt c6 protein sequences.
- Convert the protein sequences to nucleotide sequences, and use these results to implement, test, and evaluate the algorithms.

## Future Work

Investigate the role of motif in the protein-protein interaction of PSI PsaF and Cyt c6.