2003

Identification and functional analysis of thylakoid membrane proteome

Yingchun Wang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Genetics Commons, and the Molecular Biology Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/751

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Identification and functional analysis of thylakoid membrane proteome

by

Yingchun Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Majors: Genetics; Bioinformatics and Computational Biology

Program of Study Committee:
Parag R. Chitnis, Co-major Professor
Suresh C. Kothari, Co-major Professor
Steven R. Rodermel
Alan M. Myers
Xiaoqiu Huang

Iowa State University

Ames, Iowa

2003

Copyright © Yingchun Wang, 2003. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Yingchun Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program
TABLE OF CONTENTS

ABSTRACT v

CHAPTER 1. GENERAL INTRODUCTION 1

CHAPTER 2. PROTEOMICS OF MEMBRANE PROTEINS 8

CHAPTER 3. PROTEOMIC STUDY OF THE PERIPHERAL PROTEINS FROM THYLAKOID MEMBRANES OF THE CYANOBACTERIUM SYNECHOCYSTIS SP. PCC6803 50

CHAPTER 4. THE PROTEOME OF CYANOBACTERIAL PHOTOSYNTHETIC MEMBRANES: IDENTIFICATION OF NOVEL PERIPHERAL AND INTEGRAL MEMBRANE PROTEINS 70

CHAPTER 5. A SYSTEMATIC AND DETAILED FUNCTIONAL STUDY OF HYPOTHETICAL THYLAKOID PROTEINS FROM SYNECHOCYSTIS SP. PCC 6803 107

CHAPTER 6. COMPUTATIONAL OPPORTUNITIES IN PROTEOMICS RESEARCH: AUTOMATION IN DATA PROCESSING AND ANALYSIS 140

CHAPTER 7. THE MEMBRANE PROTEOME DATABASE 151

CHAPTER 8. GENERAL CONCLUSIONS 160

APPENDIX A. PROTEINS OF THE CYANOBACTERIAL PHOTOSYSTEM I 162
APPENDIX B.  THERMODYNAMICS OF ELECTRON TRANSFER IN OXYGENIC PHOTOSYNTHETIC REACTION CENTERS: A PULSED PHOTOACOUSTIC STUDY OF ELECTRON TRANSFER IN PHOTOSYSTEM I REVEALS A SIMILARITY TO BACTERIAL REACTION CENTERS IN BOTH VOLUME CHANGE AND ENTROPY 163

APPENDIX C.  THE SOURCE CODE FOR APPLICATION PROGRAMS IN THE DATABASE 164

ACKNOWLEDGMENTS 187
ABSTRACT

Membrane proteins play crucial roles in many biological functions. Identities and functions of most membrane proteins remain to be revealed because of their insolubility. New technological breakthroughs in proteomics together with the availability of genomic sequence information make it possible to study functions of membrane proteins on a genome-wide scale. We used a multidisciplinary approach combining biochemistry, genetics, proteomics and bioinformatics to study the functions of the thylakoid proteome of *Synechocystis* sp. PCC6803. The objectives of the dissertation are: (1) to identify the thylakoid membrane proteome using 2-DE and mass spectrometry coupled with database search, (2) to study the function of novel thylakoid membrane proteins using mutagenetic analysis, (3) to construct a database to facilitate the proteome study.

The thylakoid membrane proteins were separated into peripheral and integral fractions and resolved on 2-D gels with different pH range. The protein spots in the 2-D gels were subjected to peptide mass fingerprinting analysis, and totally 390 out of 558 analyzed spots were identified as protein products of 128 individual genes, of which 38 gene encode hypothetical proteins with unknown function. To study the function of the hypothetical proteins, we inactivated a set of genes, and 10 knockout mutants were obtained. The growth analysis for the mutant cells revealed that only one mutant (H1) which has a deletion on the ORF slr0110, showed conditional growth phenotype. Detailed analysis indicated that the H1 mutant is sensitive to both glucose and light. Chlorophyll fluorescence induction analysis indicated that the PQ pool of H1 is more reduced than that of wild type strain, which is caused by the electron flows from PS II and NADH dehydrogenase. We believe that it is the over-reduction of the PQ pool that makes the H1 mutant sensitive to glucose and light. The 2-D analysis of the expression profiles of the thylakoid membrane proteins of wild type and H1 mutant strains found that the plasmid psbG2 gene, which is only expressed in the photomixotrophic sensitive psbG1 deletion mutant, was also expressed in the H1 mutant. Therefore, it is highly possible that the function of the hypothetical encoded by ORF slr0110 is related with psbG1.
The ID and the structural and functional information of the identified proteins as well as the 2-D reference maps were included in a web-based relational database for thylakoid membrane proteins. The database was constructed with MySQL, and the application programs were developed with SQL, PERL, JAVASCRIPT and HTML. Users can search the information of identified proteins and compare their own identified proteins with the identified proteins in the database. A manager interface is also provided for the routine maintenance of the database.
CHAPTER 1. GENERAL INTRODUCTION

I. PROTEOME AND PROTEOMICS

Proteome is the protein complement expressed by the genome of an organism (1); it is much more complex than the genome of the organism because of the diversity in its structure, function and intracellular localization (2). To reduce the complexity, a proteome can be fractionated into sub-proteome according to structural and functional relationships, such as chloroplast proteome, mitochondrial proteome, phosphoproteome, etc. and studied in greater details. Proteomics involves the study of all protein forms expressed within an organism as a function of time, development, physiological state, and external factors. A typical procedure of proteomic analysis includes protein preparation and identification. Proteins can be prepared and fractionated (subproteome) and followed by separation using two-dimensional electrophoresis (2-DE), which has been set up as a routine tool for proteomic analysis because of its capacity of resolving thousands of proteins in a single 2-D gel (3). The resolved protein spots in 2-D gels can be analyzed by MALDI-TOF mass spectrometry and the resulting peptide mass fingerprinting can be used to search sequence databases for protein identification (4,5). A proteome contains many hypothetical proteins with unknown functions (6). Functional analysis of the hypothetical proteins is a difficult and important task in proteomics. Computational programs can be used for functional analysis by predicting topology and functional profile of proteins whereas mutagenetic analysis can be used to study functions of proteins directly (7-9). All structural and functional information obtained during proteomics analysis as well as the reference 2-D maps can be stored in databases that can help to facilitate later proteomic analysis and interlaboratory data exchanges.

II. MEMBRANE PROTEINS

Cellular membranes not only surround and compartmentalize cells, but also conduct many important biological activities. Typically, membranes consist of a lipid bilayer and associated proteins. It has been estimated that about 30% percent ORFs in an organism encode proteins with transmembrane helices (10,11). Membrane proteins can be classified as peripheral and integral proteins according to the type of their association with lipid bilayer.
Integral or intrinsic proteins span lipid bilayer one or more times with transmembrane helices that typically consist of 15-25 hydrophobic amino acid residues (12,13). Transporters, receptors, channel proteins as well as some of the photosystem proteins such as PsaAB, PsaF and the respiratory electron transport chain proteins such as cytochrome b are integral membrane proteins. Peripheral proteins do not cross the membrane and associate with lipid bilayer covalently or noncovalently. The covalent association can be achieved either through a fatty acid or polyisoprenyl chain (14), or through a glycolipid anchor, which is typically found in glycosyl phosphatidylinositol (GPI) anchored proteins (15,16). The noncovalent association is mediated by a variety of interactions, such as hydrophobic interaction, ionic interaction or hydrogen bonding.

Membrane proteins play important roles in cell functions. In all cells, proteins on plasma membrane may act as signal transduction, ion or glucose transporter. In plant or cyanobacteria, proteins on thylakoid membrane may take part in photosynthesis, such as photosystem I (PSI) and photosystem II (PSII) proteins. Some proteins are involved in respiratory pathway such as cytochrome C oxidase. Many more membrane proteins are predicted from the genome sequence, but the functions of these proteins are still unknown.

However, membrane proteins, integral membrane proteins in particular, are underrepresented in current proteomic databases because structural and functional information is scarce for these proteins due to their amphipathic nature and hydrophobicity.

### III. FRACTIONATION AND SOLUBILIZATION OF MEMBRANE PROTEINS

Peripheral proteins can be separated from lipid bilayers using reagents that can disrupt the association between them. High concentrated urea (8 M) and high pH sodium carbonate (pH 11) can be used to disrupt the hydrogen bonding and ionic interaction respectively (17-19). Covalent association can be interrupted by enzymes that can specifically cut the covalent bonds between proteins and lipid bilayer (20). The separated peripheral proteins are generally easy to be solubilized by the solutions prevailing for 2-DE. Integral proteins can be solubilized by the solutions containing chaotropes and detergents. SDS is the only detergent that can solubilize almost any integral proteins, but SDS is negatively charged and cannot be used for 2-DE, which requires electrically neutral
conditions with low salt concentration. Non-ionic detergents such as Triton X-100 or zwitterionic detergents such as sulfobetaines (SB 3-10) are often included for solubilizing integral proteins in typical 2-DE procedures. However, no detergents that has been proven as powerful as SDS to solubilize all integral proteins, a detergent that is good for one protein may not solubilize the other one (19,21-23). Therefore, solubilization of integral membrane proteins remains to be a major challenge in analysis of the membrane proteome.

**IV. MEMBRANES IN SYNECHOCYSTIS SP. PCC6803 AND THE FUNCTION OF THYLAKOIDS**

*Synechocystis* sp. PCC6803 is a unicellular photosynthetic cyanobacterium containing two types of membranes: cytoplasmic membranes and thylakoid membranes. The structure of *Synechocystis* cell is similar to that of chloroplasts in higher plants. The cytoplasmic membranes forms the inner boundary of the periplasmic space and is known to contain proteins typically associated with respiratory electron transport, such as NADPH dehydrogenase (NDH-1), cytochrome b6f, and terminal oxidase (24). The thylakoid membrane contains proteins of photosynthetic and respiratory electron transport chain as well as of many other biological processes (25,26) (Fig. 1). The photosynthetic electron transport chain includes PSI and PSII proteins, while the respiratory chain includes NDH-1, succinate dehydrogenase (SDH) and cytochrome ss3-type terminal oxidase (27). In recent proteomic analyses, the thylakoid proteins have been identified in different groups and many hypothetical proteins was observed (4,32-34). The function of the hypothetical proteins can be studied by mutagenetic analysis.
V. ORGANIZATION OF THE DISSERTATION

The dissertation includes five research chapters and one literature review chapter on membrane proteomics. In order to globally study the thylakoid proteins, we tried to resolve the thylakoid proteins into 2-D gels. This work has been divided into two manuscripts. In the first half of the work, peripheral proteins were solubilized directly and separated by 2-DE, 78 protein spots were identified as expression products of 51 individual genes by peptide mass fingerprinting. This part of work has been summarized in the chapter “Proteomic study of the peripheral proteins from thylakoid membranes of the cyanobacterium Synechocystis sp. PCC 6803” and published in the Electrophoresis 2000, 21, 1746-1754. In the second half of the work, thylakoid proteins were fractionated into peripheral and integral fractions before separating by narrow pH range 2-DE, more peripheral and integral proteins including 38 hypothetical proteins have been identified. This part of work is covered by the chapter “The proteome of cyanobacterial photosynthetic membranes: Identification of novel peripheral and integral membrane proteins” and will be submitted to the Proteomics. To identify functions of the thylakoid proteins identified in the previous studies, genes for ten hypothetical proteins were inactivated. Only one mutant that has a deletion of the ORF slr0110 shows obvious growth phenotype. Detailed functional analysis of this mutant is included in the chapter “Functional study of hypothetical thylakoid proteins from Synechocystis sp. PCC 6803”, and will be submitted to The Journal of Biological Chemistry. In the process of proteome analysis, computational programs have been developed to automate labor-intensive and time-consuming procedures. The chapter “Computational Opportunities in Proteomics Research: Automation in data processing and analysis” includes the work of automation of database search and topology prediction of membrane proteins, and the work was published in the High Performance Computing Conference (HPC 2002), December 16-19 inBanglore, India. The last chapter “The thylakoid membrane proteome database” includes the work of constructing a web-based relational database for the identified thylakoid membrane proteins. The appendices contain unrelated studies and supplementary material. Appendix A is a paper published in Biochimica et Biophysica Acta, 1507, 32-40, 2001. Appendix B is a paper published in Biochemistry, 40, 7109-7116, 2001. Appendix C includes the source code for application programs in the database.
REFERENCE


CHAPTER 2. PROTEOMICS OF MEMBRANE PROTEINS

1. Introduction to proteome and proteomics

1.1. Proteome and proteomics

Proteome is a term coined in 1994 by Mark Wilkins as the protein complements expressed by the genome of an organism (1). Proteomes are more complex than genomes because of dynamic changes due to environmental and internal cues. For examples, the proteome of plant leaf cells is different from that of root cells from the same plant even though all vegetative cells in that plant will have the same genome. In addition, the proteome would change if the plant suffered from dry condition for a period time. Proteomics is the study of all protein forms expressed within an organism as a function of time, age, physiological state, external factors, etc. Therefore, proteomics characterizes the behavior of a system rather than the behavior of a single component. Proteomic study includes at least two steps: genome-wide identification and global functional study of putative proteins. To date, the vast majority of proteomic research conducted includes at least one of the two steps. Therefore, we will discuss the proteome characterization and functional study in detail.

When the tools for large-scale protein separation and identification became available in the 1990s, high throughput proteomic studies became a reality. 2-DE in conjunction with mass spectrometry provided the most powerful tool to date for proteome analysis. The procedure of protein separation and identification using 2-DE and mass spectrometry typically contains the following steps: Protein isolation from cell or tissue; 2-DE to separate proteins according to $pI$ and molecular weight; protein spots detection and image analysis; In gel digestion with appropriate protease to generate small peptide fragments; MALDI-TOF mass spectrometry to generate peptide mass fingerprinting (PMF). The PMF data will be used for database search against the total genomic sequence to match the query data with a putative ORF (2). Finally, functional study of proteins at global level can be conducted.

The genomes are usually large and encode many more proteins than that could be resolved by any 2-DE gels (3), especially for animal or higher plant genomes. For example, the human genome is predicted to contain about 30,000-40,000 genes, and codes for a far larger proteome due to alternative splicing (4). A typical 2-DE gel can resolve no more than a
few thousand proteins, which cannot meet the requirement of resolving the whole proteome. One approach to address this issue is to fractionate the whole proteome into sub-proteomes according structural or physiological relationships prior to protein separation and identification. For example, proteins from organelles such as chloroplast or mitochondria can be fractionated as individual sub-proteomes, and proteins related with a specific metabolic or physiological pathway can also be fractionated as a sub-proteome such as photosynthesis proteome or signal transduction proteome. Here, we will concentrate on the research conducted on membrane proteome.

1.2. Membrane and membrane proteins Membrane is an important part of a cell, it provides a barrier against extracellular environment, compartmentalizes cells and performs many important biological activities. Typically, membranes consist of a lipid bilayer and proteins that are embedded or peripherally associated with the bilayer. It was estimated that about 30% percent of ORFs in an organism encode membrane proteins (5,6). However, membrane proteins especially integral membrane proteins are underrepresented in current proteomic database. Later in this chapter, we will discuss in detail the problems encountered and advances made in membrane proteomics research.

Membrane proteins play important roles in cell functions. In animal cells, proteins on plasma membrane may act as signal transduction, ion or glucose transporter. In plant or cyanobacteria, proteins on thylakoid membrane may take part in photosynthesis, such as photosystem I (PSI) and photosystem II (PSII) proteins. Some proteins are involved in the respiratory pathway such as cytochrome C oxidase. Many more membrane proteins are predicted from the genome sequence, but the functions of these proteins are still unknown.

Membrane proteins associate with a lipid bilayers in several ways. Integral or intrinsic proteins span the lipid bilayer for one or more times. These proteins normally have one or more transmembrane domains that mainly consist of hydrophobic amino acid residues. Transporter proteins, channel proteins, receptor proteins as well as some of the photosystem proteins such as PsaA-B, PsaF and those respiratory chain proteins such as Cytochrome b proteins are integral proteins. Peripheral proteins are those proteins associated with membrane with various ways, but have no transmembrane domain. The association may be
variable among the different proteins. Some proteins are modified post-translationally; a fatty acid or polyisoprenyl chain may link to the polypeptide chain, which may result in physical association with membrane. The second type of association can be achieved by a glycolipid anchor, such as glycosyl phosphatidylinositol (GPI) anchored proteins (7,8). The third type of association is mediated by protein-protein interaction, i.e. a soluble protein associate with membrane by interacting with integral membrane proteins. The example is PS I proteins PsaC, D, and PsaE. All three of these proteins are soluble and associate with the membrane through interaction with other integral membrane proteins by forming a protein complex. This kind of association is very common both in prokaryotes and eukaryotes. For instance, ATP synthase β subunit is soluble alone but associate with other integral subunits to form a functional complex. If the pure thylakoid membrane is washed with chaotropic ion or urea to separate peripheral from integral fractions, ATP synthase β subunit always present in the peripheral fractions (9). Since variety kinds of association exist between proteins and membrane. The term membrane protein is tentative except for those proteins have transmembrane domains. A protein can be called membrane proteins depends on the strength of its association with membrane and the extracting power of washing solution used to remove soluble proteins.

2. Important technology used in proteomics

2.1. Two-dimensional electrophoresis

The first high resolution 2-DE map of proteins was obtained in 1975 (10). In that exciting work, as many as 1100 different protein spots from Escherichia coli were resolved in a 13 cm 2-D gel. From that time, 2-DE has been used as a standard tool for analyzing proteins from complex biological sources.

Two-dimensional electrophoresis includes the first dimension iso-electric focusing (IEF) and the second dimension SDS-PAGE. IEF separates proteins according to their pI. Proteins are amphoteric molecules bearing negative, positive or zero net charge depending on the pH of their surrounding environment. The pI of a protein is the specific pH at which the net charge of the protein is zero. Different proteins may have different pI, mostly in the range of pH 3-12 (11). During IEF, a charged protein will migrate under the force of electric field until it moves to the place where the local pH is equal to its pI. At this point, the protein will
be zero net charged and stopped here since there is no more electrical force to move it. Therefore, proteins with different pI can be separated at the places with different pH. The second dimensional SDS-PAGE separates proteins according to their molecular weight. Proteins separated by the first dimensional IEF will be transferred to the SDS-PAGE gel, where all the proteins will be bound with highly negatively charged SDS, and migrate from cathode to anode under the force of electric field. The small proteins will migrate faster than proteins with higher molecular weight.

SDS-PAGE is a trouble free technology compared to IEF. To date, the vast majority of the problems encountered in 2-DE separation of proteins were caused by IEF. At the early stage of IEF, carrier ampholytes were used to generate the pH gradients (CA-PG). Two inherited problems of CA-PG prevented its wide use in protein analysis. One is the so called cathodic drift, which is a phenomenon found during IEF that the basic end of pH gradient collapses with increasing Vh (12). The cathodic drift is mainly due to the electroendosmosis (13), and the gradient drift can cause a flattening of the pH gradient at each end, particularly above pH 9. Therefore, basic proteins cannot be well separated by IEF. This could produce adverse effect on the separation of membrane proteins, because many integral proteins are basic proteins. The other problem of CA-PG is the low reproducibility. CA are mixed polymers that are not well characterized and suffer from batch to batch variations, these variations has detrimental effects on the reproducibility of the first dimension IEF separation. As we mentioned above, the CA-PG is unstable and could drift toward the cathode during IEF, this will also largely reduce reproducibility. Furthermore, the tube gel of CA-PG is difficult to handle, it is tend to suffer from stretch or break that may also affect the reproducibility. Therefore, CA-PG is not an ideal tool for large scale proteome analysis.

The problems associated with CA-PG can be perfectly avoided by immobilized pH gradient (IPG). IPG was firstly generated by covalently linking buffering groups (immobiline) with gel matrices (14). The buffering groups (immobiline) of IPG are a set of well characterized acrylamide monomers, each linked with a single acidic or basic group. Available immobiline species make it possible to generate any narrow linear pH gradient between 3 and 10. Because of its pH gradient stability, IPG has been proven more advantageous than CA-PG. IPG has higher resolution and higher loading capacity, and has
completely overcome the problem of cathodic drift. Furthermore, IPG is commercial available and has higher reproducibility, this is the key attribute that make IPG the routine tool for proteome analysis. The two pH gradient technologies are compared in table 1.

<table>
<thead>
<tr>
<th></th>
<th>Carrier ampholyte pH gradient (CA-PG)</th>
<th>Immobilized pH gradient (IPG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathodic drift</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Sample Loading</td>
<td>Anode Loading</td>
<td>In-gel rehydration</td>
</tr>
<tr>
<td>Lading capacity</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Compatible with basic proteins</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Commercial available</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Easy to handle</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Need to handle toxic acrylamide monomers</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

2.2 Mass Spectrometry  The emergence of proteomics as a new experimental approach to perform globally functional study of proteins is partly due to the development of mass spectrometry, which largely simplified protein analysis and characterization. Mass spectrometry has been used for analyzing biomolecules since 1980s, but has not played significant roles on a large scale until 1990s (15). This is because large and polar biomolecules are difficult to be transferred into gas phase and ionized. The matrix assisted laser desoption ionization (MALDI) and electrospray ionization (ESI) are two gentle ionization methods developed in late 1980s that can ionize single or multiple charged biomolecules without fragmentation, both of them can offer picomole to femtomole sensitivity (16). Since then mass spectrometry has become a widespread analytical tool for protein analysis.
2.2.1. MALDI-TOF

MALDI was developed as an ionization method for biomolecules in 1988 by Hillenkamp and Karas (17). MALDI is typically used in conjunction with time-of-flight (TOF) mass analyzers. MALDI-TOF is good for analyzing both large molecules and small molecules, and it also has a very good reputation for analyzing heterogeneous samples. Because of its unique power of analyzing complex mixtures, MALDI has been widely used in peptide mass fingerprinting (PMF) based protein identification process. The sample preparation for MALDI-TOF is simple and easy to handle. The matrix and analytes can be dissolved in aqueous solution and mix together prior to be co-crystallized on a metal surface, the cocrystalized mixture can subsequently be excited by pulsed laser beams.

Matrices for MALDI are typically small organic molecules with absorbance at the wavelength of the laser employed. The first matrix used for MALDI is nicotinic acid (17). More new matrices have been employed in the past years. Cinnamic acid derivative matrices (sinapic acid (18) and α-cyano-4-hydroxycinnamic acid (CHCA) (19)) and 2, 5-dihydroxybenzoic acid (DHBA) (20) are most commonly used matrices in now days. The preparation of matrix has been extensively described by Shevchenko et al. (21, 22), and the peptide separation range of each matrix is listed in table 1. Different matrices can transfer different amount of energy to analytes during desorption and ionization and therefore caused different degree of fragmentation. The CHCA matrix that generally makes highest sensitivity in MALDI is used for ions required for microseconds stability while the DHBA matrix is used for milliseconds stability (15). In order to acquire high sensitivity and resolution for a compound of peptides with different masses, different matrices should be adopted for MALDI-TOF accordingly.

The detailed mechanism of ionization process in MALDI is still unknown. The signal intensity could be affected by several factors as reviewed by Mann et al. (15). First, the incorporation of the peptides into crystals could affect the signal intensity. It has been shown that the peptides are distributed throughout the crystals rather than confined to their surface (22). Second, the ability of the peptides to capture or retain a proton during the desorption process can affect the signal intensity. Several other factors could also affect the signal intensity. For example, peptides with arginine at C-terminal can produce higher signal
intensity than peptides with lysine at c-terminal (23). Therefore, it is not practical to quantitate samples by relating them with the signal intensity, that is, the height of peaks in a MALDI-TOF spectrum. The other limitation of MALDI-TOF is that it is not good for peptides with mass range below 500 daltons, because the signal produced from these peptides are normally obscured by the signal produced from matrix related ions.

2.2.2. ESI-MS  ESI-MS has a different ionization source from MALDI. In principle, the peptide molecules in solution pass through a needle with high electrical potential and sprayed into a mist of small highly charged droplets. The dispersed droplets can be evaporated and the positively charged proteins can be released into gas phase. The protonated peptide molecules in gas phase will enter the vacuum of a mass spectrometer and be detected by mass spectrometer according to their m/z ratio (16,24).

ESI-MS has the advantage of easy to be automated. ESI-MS need a constant delivery of sample solution and is therefore easily coupled with HPLC, which is called LC-MS (25,26). LC-MS can provide more purified protein or peptide samples for MS analysis. The other advantage of ESI-MS is protein or peptide sequencing. Two ESI-MS can be coupled to form a tandem ES-MS/MS as extensively reviewed by Mann et al. (24). In principle, the protein or peptide ions of interest will be picked up by the first MS, and then fragmented in the place between the two MS/MS, usually by collisions with inert gas molecules. The collisions could create a break in the amide bonds linking the amino acids, resulting in peptide fragments with a charge retained either on C-terminal or N-terminal. The second

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinapinic acid (3, 5-dimethoxy-4-hydroxy cinnamic acid)</td>
<td>Peptides and Proteins &gt; 10,000 Da</td>
</tr>
<tr>
<td>a-cyano-4-hydroxycinnamic acid (CHCA)</td>
<td>Peptides &lt; 3,000 Da</td>
</tr>
<tr>
<td>2, 5-dihydroxybenzoic acid (2, 5-DHBA)</td>
<td>Peptides</td>
</tr>
</tbody>
</table>
MS could detect the charged fragments. Since the each amino bond in a peptide could break and form a serial of fragments with one amino acid different in length, the sequence of the peptide could be deduced based on the mass difference of these fragments. The ESI-MS sequencing is fast and sensitive compared with conventional Edman degradation (27), but has a limitation in peptide length (15 amino acid residues) and cannot differentiate isoleucine from leucine. Therefore, it has not been widespread adopted as a substitute of Edman degradation sequencing. Furthermore, ESI-MS is not very tolerant of salts, a good sample for ESI-MS should contain less than 1 mM ionic salts or 0.01% ionic detergents (24).

2.2.3. SELDI SELDI-MS is a newly developed MS technology with simplified sample extraction for protein analysis (28). In principle, protein mixture from cells or tissues could be fixed by interacting with immobilized proteins on the surface of a protein chips. The interaction could be either chemical (such as ionic, hydrophobic and hydrophilic and etc.) or biochemical (such as antibody, receptor, DNA and etc.) (29). The un-fixed proteins could be removed by simple washing with a solution. The fixed proteins could desorp and ionize by the shot of laser beam. SELDI-MS could detect intact proteins with high sensitivity and is therefore rapidly used for finding protein markers for medical purpose. One of the biggest problems of SELDI-MS is that can detect only a limited number of proteins because of the limitation of available protein chips. The other problem is that SELDI-MS can only be used to detect the mass of proteins, thus SELDI-MS cannot be used for protein identification based on peptide mass fingerprinting. Therefore, there use of SELDI-MS is currently limited to medical research such as protein profiling and early disease detection.

2.3 Protein detection technology The resolved protein spots on 2-DE gels can be visualized using staining technology. Three staining technologies have been widely employed as laboratory routines. These are silver staining, fluorescence staining and Coomassie blue staining. The advantage and drawback of each staining methods have been described in detail (30,31), as compared in Table 3.

In addition to the staining methods discussed above, negative staining such as zinc-imidazole reverse staining (32,33) and immuno-blot are also been used to detect proteins on 2-DE gels. The stained image of 2-DE gels can be analyzed by commercial available
software Melanie or PDQuest (Bio-RAD). Both of the software can perform quality and quantity as well as statistic analysis.

<table>
<thead>
<tr>
<th>Staining</th>
<th>Detection Limits</th>
<th>Linear Dynamic Range</th>
<th>Compatible with mass spectrometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver Staining</td>
<td>4-8 ng</td>
<td>8-60 ng</td>
<td>No</td>
</tr>
<tr>
<td>Fluorescence Staining</td>
<td>1-2 ng</td>
<td>1-1000 ng</td>
<td>Yes</td>
</tr>
<tr>
<td>Colloidal Coomassie Blue Staining (G-250)</td>
<td>8-16 ng</td>
<td>30-250 ng</td>
<td>Yes</td>
</tr>
</tbody>
</table>

2.4 Current membrane protein databases Protein databases are useful tools for the structural and functional study of proteins. One of the earliest protein databases is The Protein Data Bank (PDB). The PDB was established in 1971 and includes information of crystal structures of proteins. A similar database can also be reached at Stephen White Laboratory at UC Irvine; this database also includes crystal structures of integral membrane proteins. Today, with fast technological development in genomics and proteomics, together with more available powerful computational tools, more protein databases that include information other than three-dimensional (3D) structures of proteins have been established (Table 4). The Transporter Database at UCSD classifies all known membrane transporters from 39 organisms according to their function. The SWISS-2DPAGE provides reference maps of proteomes for human, mouse, Arabidopsis, E. coli and yeast, but little structural and functional information is provided. The Arabidopsis Membrane Protein Library (AMPL) classifies membrane proteins into 13 categories such as ABC transporters, cell wall biosynthesis, ion channels and etc. Secondary structural information such as transmembrane helix and its orientation is also included in this database. Beside these, the proteome project in Cyano2Dbase also provides a reference map of the thylakoid proteome that mainly
includes PSI and PSII proteins of *Synechocystis*. Despite more and more available protein database, there is no specific database for membrane proteome. Therefore, an annotated 2-DE protein database specific for membrane proteins should be constructed as described previously for other proteins (34-38).

**Table 4. Current databases for structural and functional information of proteins**

<table>
<thead>
<tr>
<th>Type</th>
<th>Database Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWISS-PROT</td>
<td>Annotated protein sequence database</td>
<td><a href="http://us.expasy.org/sprot/">http://us.expasy.org/sprot/</a></td>
</tr>
<tr>
<td>Protein 3D structure and associated topics databases</td>
<td>PDB</td>
<td>Protein Data Bank for 3D structure</td>
<td><a href="http://www.rcsb.org/pdb/">http://www.rcsb.org/pdb/</a></td>
</tr>
<tr>
<td></td>
<td>TOPS</td>
<td>Protein topology atlas</td>
<td><a href="http://www3.ebi.ac.uk/tops">http://www3.ebi.ac.uk/tops</a></td>
</tr>
<tr>
<td></td>
<td>SCOP</td>
<td>Structural classification of proteins</td>
<td><a href="http://scop.mrc-lmb.cam.ac.uk/scop/">http://scop.mrc-lmb.cam.ac.uk/scop/</a></td>
</tr>
<tr>
<td></td>
<td>BMM Domain Server</td>
<td>Biomolecular Modelling Laboratory (ICRF) protein domain server</td>
<td><a href="http://www.bmm.icnet.uk/~domains/">http://www.bmm.icnet.uk/~domains/</a></td>
</tr>
<tr>
<td>Protein motif and domain databases</td>
<td>PROSITE</td>
<td>Dictionary of protein sites and patterns</td>
<td><a href="http://us.expasy.org/prosite/">http://us.expasy.org/prosite/</a></td>
</tr>
<tr>
<td></td>
<td>InterPro</td>
<td>Integrated Resources of Proteins Domains and Functional Sites</td>
<td><a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a></td>
</tr>
<tr>
<td></td>
<td>Pfam</td>
<td>Protein families db (HMM derived)</td>
<td><a href="http://pfam.wustl.edu/">http://pfam.wustl.edu/</a></td>
</tr>
<tr>
<td></td>
<td>PRINTS</td>
<td>Protein Motif fingerprint database</td>
<td><a href="http://bioinf.man.ac.uk/dbbrowser/PRINTS/">http://bioinf.man.ac.uk/dbbrowser/PRINTS/</a></td>
</tr>
<tr>
<td></td>
<td>SBASE</td>
<td>Proteins domains database</td>
<td><a href="http://www3.icgeb.trieste.it/~sbasesrv/">http://www3.icgeb.trieste.it/~sbasesrv/</a></td>
</tr>
</tbody>
</table>
### Proteome databases for multi-species

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIENA-2DPAGE</td>
<td>2-DE PAGE database for Human, <em>Chlamydia trachomatis</em>, <em>Caenorhabditis elegans</em></td>
<td><a href="http://www.bio-mol.unisi.it/2d/2d.html">http://www.bio-mol.unisi.it/2d/2d.html</a></td>
</tr>
</tbody>
</table>

### Proteome databases for a specific organism

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyano2Dbase</td>
<td>Proteome database for <em>Synechocystis</em> sp. PCC6803</td>
<td><a href="http://www.kazusa.or.jp/cyanobio/Synechocystis/cyano2D/index.html">http://www.kazusa.or.jp/cyanobio/Synechocystis/cyano2D/index.html</a></td>
</tr>
<tr>
<td>YPD-Yeast Protein Map</td>
<td>Proteome database for yeast</td>
<td><a href="http://www.ibgc.u-bordeaux2.fr/YPM/">http://www.ibgc.u-bordeaux2.fr/YPM/</a></td>
</tr>
<tr>
<td>The Arabidopsis mitochondrial proteome project</td>
<td>Proteome database for Arabidopsis mitochondrial proteins.</td>
<td><a href="http://www.gartenbau.uni-hannover.de/genetik/AMPP">http://www.gartenbau.uni-hannover.de/genetik/AMPP</a></td>
</tr>
</tbody>
</table>

### Membrane protein databases

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
</table>

### 3. Solubilization of membrane proteins

Solubilization is the crucial step for 2-DE separation of membrane proteins. Membrane proteins are usually hard to be solubilized into aqueous solutions and most of the problems encountered during 2-DE separation are originated from the insolubility. The solubility of membrane proteins affected by many factors, we will discuss them one by one in the following paragraphs.

#### 3.1. Hydrophobicity

Hydrophobicity is an internal attribute of a protein and its value depends on the protein’s amino acid composition. The hydrophobicity of a protein can be
evaluated using the GRAVY score (39) or the number of trans-membrane domains (TM). GRAVY score is the average hydrophobicity index of a protein. The GRAVY value of a protein is calculated by assigning a hydropathy value to the side chain of each amino acid residue. For example, Serine was assigned to -0.8, Cysteine was assigned to 2.5. The higher value means the amino acid is more hydrophobic. The most hydrophilic amino acid is Arginine, which have a value of -4.5; the most hydrophobic amino acid is Isoleucine, which has a value of 4.5 (39). The GRAVY score of a protein is the arithmetic mean value of all of its amino acid residues. GRAVY values of 10260 proteins from three organism *Escherichia coli, Bacillus subtilis* and *Saccharomyces cerevisiae* have been calculated, the highly hydrophobic protein has the highest GRAVY value of 1.7 while the highly hydrophilic proteins has the lowest GRAVY value of -2.2 (40). However, the highest GRAVY values calculated from proteins identified on 2-DE gels from these three organisms are 0.1, 0.15 and 0.3 respectively. Therefore, many highly hydrophobic proteins are lost during 2-DE process either from poor protein solubilization during sample preparation, precipitation during IEF or inefficient transfer from first to second dimension (40,41). The higher GRAVY score may mean the protein is more hydrophobic. But actually, it is not always true, because some large membrane proteins may have lower GRAVY score than small soluble proteins as large proteins tent to have more loop regions which predominately consist of hydrophilic amino acid residues. Therefore, it is possible that a more hydrophobic protein may have lower GRAVY values than a hydrophilic protein. In such a case, the number of TMs should be considered in order to accurately evaluate the hydrophobicity of the protein. More TMs may mean a protein is more hydrophobic. To date, the protein that has the most number of TMs identified on 2-DE gels is the human red blood cell ghost protein band III. This protein has 12 TMs and can only be resolved on 2-DE using new synthesized detergents (42,43). Unfortunately, many membrane proteins with more than one TM segment remain to be resolved by 2-DE.

**3.2. Protein size and pI** Most proteins from a proteome are predicted to be within the mass range 10-100 kD and the pI range 4-7 (11,44). High molecular weight proteins are hard to be resolved into 2-D gels; they tend to precipitate during IEF. The precipitated proteins remain in first dimension gel and cannot be detected in the second dimension gel. The mass
range that can be covered by 2-DE is typically from 8 kD to 100 kD (45), even though higher molecular weight proteins could be recovered sometimes (40,46). pI is the other factor that affects the solubility of proteins for 2-DE. It was estimated that about 20% ORFs of an organism encoding proteins with pI higher than 10.0 (47), and most of the integral membrane proteins have an alkaline pI (40). Currently, the pH range of most of the widely used IPGs is between 3 to 10, which is not enough to resolve proteins with pI more than 10. Although extended pH gradients such as pH 6-11 are also available commercially, using this kind of IPG strips bring some new problems. Horizontal streaking is very common in 2-DE gels with pH range 6-11 because of endosmosis effects (48). Fortunately, more efforts have been spend on 2-DE separation of alkaline proteins recently. In one report, IPG strips with high alkaline pH gradient were pre-washed with ascorbic acid before IEF and followed by increased temperature for IEF, alkaline proteins can be well resolved on 2-DE gels and horizontal streaking can be totally abolished (49). Following this method, we also get a good resolution for alkaline proteins using IPG strips with pH range 6-11. In the other report, Glycerol and isopropanol were added to the IEF solution to alleviate the problems caused by endosmosis, the 2-DE pattern for alkaline proteins has been largely improved (50). Furthermore, most of the proteins have the lowest solubility at pH equal to their pI. Therefore, some proteins especially membrane proteins tend to precipitate at the places in the immobilized pH gradients where the local pH is close to their pI during IEF, and hence could not migrate to the second dimension (51). The precipitated proteins could form visible bands in the IPG strips.

3.3. Chaotropes

Chaotropes are used to disrupt intra- or inter-protein hydrogen bonding. The disruption of hydrogen bonding can not only lead to protein unfolding and denaturation, but also disrupt the structure of water molecules and thereby relax the constraint on hydrophobic amino acid side chains (52). Urea is a neutral chaotrope; high concentration of urea can increase the solubility of many membrane proteins. Urea can unfold membrane proteins and expose their hydrophobic core to aqueous solution by disrupting the hydrogen bonding that keep proteins in native state (53,54). Protein solubility could be increased through using urea in conjunction with thiourea (55,56). Thiourea is hard to dissolve in water and high concentration of urea can increase its solubility. Typically, 2 M
thiourea is used together with 7 mM urea. Higher concentration of thiourea in IEF solution (> 2 M) could lead to loss of protein resolution, the possible reason is that high concentration of thiourea may inhibit the SDS-protein binding during the protein transferring from the first dimension IPG strip to the second dimension SDS-PAGE (57).

### 3.4. Detergents

All of the research used detergents have a common character, that is, all of them have a polar head and a hydrophobic tail. Solubilization of membrane proteins with the help of detergents has been clarified in mechanism. The hydrophobic groups of detergents can bind to proteins by hydrophobic interaction, that is, the hydrophobic tail can infiltrate into lipid bilayers and bind to membrane proteins while the polar head interact with surrounding water. In this way, proteins would be solubilized in detergent-bound form. Some newly synthesized detergents may have two or more polar heads (58), but this kind of detergent is generally not as efficient as those that have similar hydrophobic tails but only one polar head. The possible reason may be due to the hydrophilic interactions between polar heads and hence weak the interaction between polar heads and aqueous solution. The length of the hydrophobic tail also has effects on the efficacy. Too short tail will not be hydrophobic enough to bind proteins, too long tail will be too hydrophobic to solubilize. A tail with 14-16 carbons seems to be the optimum length (58,59). The most strong and widely used detergent used for solubilizing proteins is SDS; it has a long hydrophobic tail, which can interact with and bind to proteins. The polar head of SDS is $\text{SO}_3^2$, which is double negatively charged and can interact with water. The ratio of binding is 1.4g SDS per g protein, which can approximately counted as one SDS molecule per every two amino acids (60-62). SDS can solubilize almost any proteins. Since SDS is high negatively charged, it could not be used for 2-DE because IEF requires low salt and electrically neutral detergent conditions. Although some protocol suggest that SDS can be used to extract proteins from membrane followed by extensive dilution with non-ionic or zwitterionic detergents (54,63,64), it is not applicable when high load samples for IEF are required. The original detergents used for 2-DE purpose are NP-40 and Triton X-100 (10,65), but these detergents combined with urea were not efficient to solubilize highly hydrophobic proteins. In recent years, CHAPS have been proven to be a more efficient zwitterionic detergent. Routinely, 4% CHAPS is used in conjunction with 7M urea. However, higher concentration
of urea and thiourea were observed to decrease the solubilizing power of CHAPS (43,60). It was assumed that high concentration of urea will denature proteins and expose the hydrophobic core to aqueous solution; thereby more sites will be available for protein aggregation and the solubility will be decreased. In order to solve this problem, two solutions were suggested to use for different proteins (56). One solution contain high concentration of urea (7M), and CHAPS (4%), the other solution contains low concentration of urea (5M) and CHAPS (2%), and also contains 2% sulfobetaines 3-10 (SB 3-10) with a long linear hydrophobic tail. The latter has been proven to be more efficient than CHAPS in high concentration of urea. SB3-10 also has a drawback, it has low solubility in high concentration of urea, so when perform IEF at 20 °C in a solution containing urea and SB 3-10, detergent crystallization is always a big problem. An increase of the running temperature from 20 °C to 25 °C can minimize this problem. Some new synthesized detergent have also been applied to 2-DE, such as amidosulfobetaines (ASB-n), alkylaryl amidosulfobetaines (Cnφ). The most efficient detergents are ASB-14 and C8φ, each one can increase the solubility of some hydrophobic proteins (43,58,59,66).

The efficiency of several commercial available nonionic or zwitterionic detergents has been recently tested based on 2-DE solubilization of band III protein of human red blood cell ghosts. The detergents bearing sugar-derived polar head such as α (or β)-dodecyl maltoside are more efficient than detergents bearing oligoethylene glycol polar head such as Decaethylene glycol mono oleyl ether (Brij 96) or Triton X-100 (67). Different proteins may have different solubility in different detergents. As an example, ASB 14 has a stronger solubilizing power than C7BzO for band III while C7BzO is clearly more efficient for plant membrane proteins. The condition could be more complicated if considering together with chaotrope. For example, CHAPS has been known more efficient than Triton X-100 when urea alone is used as chaotrope while Triton X-100 is more powerful than CHAPS if urea-thiourea mixture is used as chaotrope for solubilizing membrane proteins (68).

The other factor that could affect the efficiency of detergents is sample treatment. Best solubilization can only be achieved through using different detergents according to different sample treatment (66). For example, different detergents are required for delipidated
or non-delipidated sample for best solubilization, as shown by Santoni et al. (53). The insoluble fractions of plasma membrane of *Arabidopsis thaliana* after extracting with a solution containing Triton x-100 were either delipidated or non-delipidated. If the insoluble fractions were delipidated, then using an extraction buffer containing detergent §1 will get the best solubilization of integral proteins; if the insoluble fractions were directly extracted with an extraction buffer without delipidation step, then an extraction buffer containing ASB14 will provide the best results (53).

In general, there is no such a universal detergent that can be used for solubilizing all membrane proteins. Delicate detergents should be used for different membrane proteins. The optimal choice of a detergent still remains largely empirical. As an alternative, a mixture of several detergents may meet the requirement of solubilizing various membrane proteins.

### 3.5. Reducing agents

Reducing agents help to solubilize proteins by disrupting disulfide bonds existing intra- or inter-proteins. The traditional reducing agents used are dithiothreitol (DTT) or dithioerythritol (DTE) (69). However, DTT is not a trouble free reducing agent for 2-DE. During IEF, the free thiol groups of DTT are able to obtain a charge and migrate out of the pH gradient. Therefore, the reducing condition of IEF could be lost and free thiol groups of proteins may form the disulfide bond again which eventually lead to the precipitation of proteins. Furthermore, two-step equilibration is often required for vertical second dimension system before transferring proteins from the first to the second dimension if using DTT as the reducing agent, because residual DTT could lead to point streaking and other silver staining artifacts on the second dimension gel. Iodoacetamide can be used in the equilibration solution in order to alkylate thiol group of DTT and minimize the artifacts on the resulting gel (70,71). In recent years, a more powerful and efficient reducing agent has been introduced for interrupting disulfide bonds (72). The non-thiol containing reducing agent tributyl phosphine (TBP) has proven to be more advantageous than DTT (73). TBP has no charged group; it does not migrate in IPG during IEF. Therefore, TBP can maintain the reducing conditions for the entire IEF course. The increased solubility provided by the maintenance of reducing conditions gives improved focusing and decreased horizontal streaking on the subsequent second-dimension gel (73). For example, the four isoforms of wool keratin with closely related mass and pI could be well resolved on 2-DE gels if using
TBP as reducing agent, but can not get good resolution if using DTT instead of TBP (73). As TBP has no thiol group; it does not react with thiol specific alkylating reagents. Therefore, it is feasible to run IEF with iodoacetamide in IEF sample solution to alkylate free thiol group of cysteine and thus prevent their reoxidation (59,73), and two-step equilibration is not necessary by incorporating reduction and alkylation in a single step. The only problem for TBP is the price and not easy to handle. Concentrated TBP is volatile, toxic and highly flammable. Shipment of TBP is considered dangerous. Therefore, some company such as Bio-Rad provide TBP in a diluted form with a concentration of 20 mM TBP in 2-propanol.

3.6. Carrier ampholytes  A mixture of carrier ampholytes not only improve separations by producing more uniform conductivity across the pH gradient during IEF, but also can increase solubility of proteins by minimizing protein aggregation due to charge-charge interactions. It has been observed that carrier ampholytes can reduce protein-matrix hydrophobic interactions and circumvent the detrimental effects caused by salt boundaries (74). Some proteins need stringent salt requirement to maintain their solubility, a certain amount of salt should be present in the sample solution for IEF. This will cause some problems. First, salt ions will affect IEF by limiting the voltage that can be achieved without producing high current, which will increase the time required for IEF. Second, salt ions will migrate to the anode or cathode, thus cause the precipitation of some proteins that require salts to maintain their solubility. Carrier ampholytes can be used as a substitute of salts to overcome these problems. Typical concentration of Carrier ampholytes is between 0.5% and 2% depending on the amount of proteins loaded; high concentration will slow down IEF since they also carry charges.

3.7. Lipids  The hydrophobic interaction between proteins and lipids is one of the major factors that cause the insolubility of some integral membrane proteins. Lipid not only form complex with proteins, but also form complex with detergents, this may explain why some detergent works well for delipidated samples and some for samples without delipidation (60). Furthermore, large amount of lipid in samples could block gel pores of IPG and cause smears and horizontal streaking on the subsequent second dimension gel. Organic solvents could be used to remove lipids. Several organic solvents have been tried to remove lipid from protein samples (75-77), the one that provided best results for 2-DE is acetone (78).
Delipidation can help to substantially improve the quality of overall 2-DE patterns, remove horizontal streaking and smearing, increase the intensity of some proteins and achieving better resolution (76-78). Furthermore, delipidation can help to resolve some highly hydrophobic proteins on 2-DE gels. As an example, the spectrins from red blood cell membrane can be resolved on 2-DE gels by delipidation (76-78). However, delipidation is not a trouble free technique that can be applied to any proteins. It has been observed that delipidation does not increase 2-DE resolution while reduce protein spot intensity (75). The other report indicated that delipidation could totally prevent resolubilization of some proteins such as band III (42). Furthermore, a severe loss of proteins may be suffered from delipidation because the organic solvent used can dissolve some proteins (54). Therefore, we must be conscious whether a delipidation step is required for a specific protein sample prior to IEF.

3.8. In-gel rehydration vs. cup loading

Sample application by in-gel rehydration is an important advance in the technology of 2-DE. The sample proteins are included in the rehydration buffer and migrate or are adsorbed to gel matrix in the process of rehydration. In-gel rehydration has been proven more advantageous than conventional cup loading for membrane proteins (12). For cup-loading, many proteins may lose at the loading point because of precipitation due to insufficient buffering power at the solution-gel interface, and the loss could be more serious for membrane proteins (54,57,79). The loss of proteins caused by cup-loading can be avoided by sample application using in-gel rehydration. For in-gel rehydration, a dispersion of membrane proteins in rehydration solution up to 450 µl can provide adequate buffering power and thereby decrease the possibility of protein aggregation. As we previously discussed, most of the membrane proteins are low abundant proteins that are normally shielded by those high abundant proteins in 2-D gels and can not be detected by normal staining methods (Western blot can detect low abundant proteins on 2-D gels). Thus it is a straightforward idea to increase the amount of proteins for 2-DE in order to identify the low abundant proteins. With the large loading volume of in-gel rehydration, it is feasible to load more proteins into a single IPG strips. Coupled with using narrow pH range IPG strips, it is possible to detect the low abundant membrane proteins on 2-D gels using conventional staining technology.
3.9. Passive rehydration vs. active rehydration  Active rehydration is a technology used for in-gel rehydration. The difference of active rehydration and passive rehydration is that low voltage (normally 50 V) is given over the entire IPG length for active rehydration while no voltage is given for passive rehydration. Active rehydration is more advantageous than passive rehydration. First, during active rehydration, proteins begin to migrate to positions where the local pH is equal to their pi, which can largely reduce the time for IEF, and decrease the possibility of precipitation. Second, the protein migration during rehydration can prevent local sample overload, which may lead to protein aggregation and precipitation. It is reported that active rehydration can help to resolve some high molecular weight membrane proteins such as spectrin α chain of red blood cell membrane (80) while passive rehydration cannot (43). Therefore, active rehydration is a better choice especially for membrane proteins which always suffered the problem of aggregation and precipitation.

3.10. Other factors that can affect solubilization of proteins for 2-DE  Alkylation prior 2-DE procedure can improve solubilization of alkaline membrane proteins. The current standard protocol for 2-DE requires sample reduction prior to IEF, followed by a reduction/alkylation step in between the first and second dimension. The protocol is far from optimal because the failure to alkylate reduced proteins prior to IEF will result in a substantial loss of alkaline proteins due to the regeneration of disulfide bridges, which will cause the formation of macroaggregates that will be trapped within the gel matrix (81). If the sample proteins is alkylated with twice the molarity of the reducing regent, then more alkaline proteins can be resolved on the second dimension gel, with the improvement of pattern quality (less streaking, rounder spots) (77,81).

Inefficient transfer from the first to second dimension is the other factor that causes the loss of membrane proteins. Efficiency of transfer can be increased by using IPG containing optimal concentration of polyacrylamide. Candiano et al. reported that by using home-made IPG containing 3.3% T matrix instead of commercial available IPG containing 4% T matrix, The number and the intensity of membrane proteins can be substantially increased, especially for high molecular weight proteins (> 100 kD) (77).
4. Fractionation of membranes and membrane proteins

Pure membrane must be obtained prior to any subsequent proteomic analysis. Cell membrane system includes plasma membrane, organelle membrane (such as thylakoid membrane, mitochondrial membrane, Golgi apparatus membrane) and nuclear membrane. Membranes can be released from cells through mechanical breaking or osmotic lysis of cells. The released membrane can be purified by sequential centrifugation to remove unbroken cells or cell debris, the carry-over cytoplasmic fraction that loosely attached to membrane can be removed by washing with low ionic strength solution. Since most of the membrane proteins are generally not very abundant, further separation of different membrane fractions is required to enrich low abundant proteins. Membrane can be fractionated with sucrose gradient centrifugation. For example, thylakoid membrane of Synechocystis can be separated and formed a dark green band on the interface between 39% and 50% sucrose layer (82). An efficient separation strategy that exploring aqueous polymer two-phase partitioning combined with sucrose density centrifugation has been reported to separate plasma membrane from thylakoid membrane in Synechocystis sp. PCC6803 (83,84). Pure plasma membrane and thylakoid membrane could be obtained without cross contamination using this method.

Sample prefractionation can not only help to enrich low abundant proteins (85), but also help to localize proteins into specific cellular compartments (86). As to membrane proteins, we can fractionate them into peripheral and integral fractions. Two approaches have been adopted in membrane protein preparation. One is sequential extraction; the other is two phase partition with detergent. In addition, organic solvent extraction and affinity binding have also been used for membrane protein extraction.

4.1. Sequential extraction  Sequential extraction is based on the idea of relative solubility of proteins in a series of solutions (66,87). Generally, the solubility of cytoplasmic proteins is higher than peripheral membrane proteins, and the latter have higher solubility than integral proteins. Therefore, pure integral membrane proteins can be obtained by removing cytoplasmic and peripheral membrane proteins through step-wise washing with solutions of different solubilizing power. One of the classic experiments for separating
membrane proteins was conducted by Molloy et al. (88). In the first step of the separation, *E. coli* cells were lysed in 40 mM Tris-base (pH 9.5) and cytoplasmic proteins were extracted. After washing with 40 mM Tris-base for two times to remove carry-over soluble proteins, the insoluble fractions from first extraction were subjected to second extraction with a solubilization solution (8 M urea, 4% w/v CHAPS, 100 mM DTT, 40 mM Tris-base and 0.5% v/v Pharmalytes 3-10, 150 U endonuclease, pH 9.5). The second extraction can remove peripheral from integral proteins. The final extraction were performed on the insoluble pellet from second extraction with an enhanced solubilization solution (5M urea, 2M thiourea, 2% w/v CHAPS, 2% w/v SB 3-10, 2 mM TBP, 40 mM Tris-base and 0.5% v/v Pharmalytes 3-10, 150 U endonuclease, pH 9.5). After this extraction, some of the hydrophobic integral membrane proteins were supposed to be extracted into the enhanced solubilization solution. The final insoluble residue was resuspended in 1% SDS and boiled for 5 minutes, then diluted 10 times with the enhanced solubilization solution. The identification results showed that membrane proteins were enriched in the final extraction. For the final residue solubilized with 1% SDS, one membrane protein was also detected.

The whole process of sequential extraction for membrane proteins could be described as: 1. Disrupt cells and prepare membrane. 2. Washing membrane with solution 1 to remove soluble contaminants. 3. Extract membrane from step 2 with solution 2 to remove peripheral proteins. 4. Extract insoluble fraction from step 3 with solution 3 to get integral proteins. 5. Use 1% SDS to extract more hydrophobic integral proteins. The solubilizing power from solution 1 to solution 3 is increased stepwise. For the solution 1 used to remove soluble contaminants, several solutions could be used other than 40 mM Tris-base. For example 0.1M sodium carbonate (pH 11) (60,87) and EDTA buffer containing 20 mM MOPS, pH 7.0, 50mM EDTA or other chaotropic salts could be used as solution I. However, since sodium carbonate removes those contaminants through the effect of its high pH, many of the peripheral proteins could also be removed. So it may not appropriate to use sodium carbonate when making peripheral proteins i s a lso a p urpose. In o ur lab, w e g enerally u se 8 M u rea instead of the solution containing 8M urea, 4% w/v CHAPS, 100mM DTT to separate peripheral proteins from integral proteins (solution 2), since the solution containing 8M urea,
4% CHAPS and 100 mM DTT will be strong enough to extract some less hydrophobic integral proteins into solutions.

4.2. Two-phase partition The nonionic detergent Triton X-114 is often used to separate hydrophilic proteins from hydrophobic proteins (66,89). Aqueous solutions containing Triton X-114 are homogeneous at low temperature near 0°C, but at high temperature such as 20 °C, two phases will be formed. The top phase is the aqueous phase; almost all of the detergent has been depleted from this phase. The low phase is the detergent phase. Concomitant with the forming of the two phases, the vast majority of the hydrophilic proteins will come into the aqueous phase and hydrophobic proteins will be only found in the detergent phase. Thus using this partition system, we can separate peripheral proteins from integral proteins. The temperature for phase partitioning could be lowered by adding glycerol to the system. The Triton X-114 fractionation has been used to fractionate integral membrane proteins from Arabidopsis thaliana leaf plasma membrane (PM) (66). Compared to the 700 spots on the 2-D gel for the total PM proteins, only 400 spots were recovered on the gel for Triton X-114 fraction. Considering the fact that 80 % proteins detected in 2-D gels corresponded nearly exclusively to peripheral proteins (90), nearly half of the peripheral proteins have been removed from the TritonX-114 fraction by the two-phase partition. The amount of five integral proteins that is specific to PM or enriched in PM when compared to the cytosolic fraction was increased while eight proteins that presented with similar abundance in PM and soluble fraction was decreased in the 2-D gel for Triton X-114 fraction, the result indicated that Triton X-114 can enrich some of proteins specific to PM and impoverish some soluble contaminants. Nevertheless, only 9 proteins were found specific to Triton X-114 fraction, and many integral proteins in the fraction were not recovered on 2-D gels. Furthermore, water channels that are one of the most abundant proteins were also not detected on the 2-D gels. Therefore, integral membrane proteins enriched in Triton X-114, could be underrepresented on 2-D gels.

4.3. Sodium carbonate washing Sodium carbonate at high pH (pH11) can release peripheral proteins from membrane by interrupting the hydrogen bonding interaction between proteins. Washing with sodium carbonate will not disrupt the membrane structure or remove any integral components (87). Sodium carbonate fractionation has been proven
more efficient in enrichment of hydrophobic proteins than Triton X-114 for the purpose of 2-DE (66). In the experiment, peripheral proteins of *Arabidopsis thaliana* leaf PM were striped out by sodium carbonate; the remaining integral fraction was then solubilized and resolved on 2-D gels. Most of the PM proteins were undetected on the 2-D gels, and more proteins are specific to integral fraction than that of Triton X-114 fraction. Several integral proteins that were undetected on 2-D gels for total PM proteins or for Triton X-114 fraction such as H+-ATPase, cellulose and water channels were present in this fraction. The result suggested that sodium carbonate fractionation could be a promising approach for 2-DE separation of integral membrane proteins.

4.4 Organic solvent extraction  Organic solvents have been shown to be able to extract hydrophobic proteins from purified chloroplast envelope membrane (91). Chloroform/methanol mixtures with different ratio have been tested for the ability to extract integral membrane proteins from PM of *Arabidopsis thaliana* leaf and the chloroplast envelope and thylakoid membrane from *Spinacia oleracea* L. (53, 91, 92). The chloroform/methanol ratio was adopted as 1:8, 2:7, 3:6, 4:5, 5:4, 6:3; 7:2; 8:1 (v/v). When extracting proteins, the organic phase contains hydrophobic proteins, and the insoluble pellets which are composed of soluble or moderate hydrophobic proteins, locate at either the interface (for chloroform/methanol solutions 6:3, 7:2, and 8:1) or at the button of centrifuge tube (for all the other chloroform/methanol solutions). The chloroform/methanol ratio 5:4 is the optimum solution for extracting integral proteins (91). The chloroform/methanol extraction is more selective in removing soluble and peripheral proteins than other sequential extraction with detergent or salt. Only hydrophobic proteins could be dissolved in organic phase (53, 92).

All of the three membrane protein extraction approaches have the advantage of fractionating peripheral and integral proteins and enriching low abundant proteins. But the underlying mechanisms are different. Serial extraction fractionate proteins by step-wise solubilizing proteins using solutions with different solubilizing power, less hydrophobic proteins will be solubilized earlier while more hydrophobic proteins will be solubilized later. Two-phase partition with Triton X-114 separates proteins according to their solubility in detergent and aqueous phases, with hydrophobic proteins in detergent and hydrophilic
proteins in aqueous phase. Organic solvent extraction is different from two-phase partition with hydrophobic proteins solubilized in organic phase while less hydrophobic or soluble proteins precipitated. Other methods for isolating membrane proteins also have been reported. For example, anion-exchange column chromatography (93) has been used to isolate hydrophobic proteins from rat liver Golgi complex; phenyl Sepharose resin affinity binding has been used to isolate mitochondria membrane proteins (94) and etc.

5. Current identified membrane proteins using 2-DE
5.1. Identified membrane proteins from prokaryotic organism and yeast

E.coli. Two proteomic studies of membrane proteins from E.coli K-12 strain W3110 have been reported. In one study, the total cell membrane fractions have been isolated, sequential extraction using a solution containing 8M urea, 4% w/v CHAPs, 100mM DTT has been used to remove peripheral proteins from integral proteins. Eleven integral proteins have been identified in the 2-D gel with pH range of 3.5-10 nonlinear gradients, six of them are outer membrane proteins (88). The other study has been performed recently on the outer membrane of E.coli. The membrane was extracted with 0.1M sodium carbonate to separate integral proteins from peripheral ones. In a 2-D gel with pH range 4-7, twenty one outer membrane proteins with size between 10-80 kD were identified, which consisted of 80% of all of the 26 predicted outer membrane proteins with pI 4-7 and size 1-80 kD in SWISS-PROT Release 37. These identified proteins including porins, transport proteins, enzymes and hypothetical proteins. Furthermore, five outer membrane-associated lipoproteins as well as three cytoplasmic associated proteins were also identified. Interestingly, all of the identified membrane proteins have negative GRAVY score, which may indicate these membrane proteins are not highly hydrophobic (95).

Corynebacterium glutamicum C. glutamicum belongs to Gram positive bacteria, there are supposed no outer membrane in this organism. However, an additional membranous structure, a layer of mycolic acids was found to covalently bind to the external site of peptidoglycan sacculus. Membrane fractions were separated from C. glutamicum wild type strain ATCC13032, cytoplasmic proteins was removed by ultracentrifugation. The membrane proteins were extracted with a lysis buffer containing 2-4% CHAPS. Membrane
proteins were resolved by 2-DE, 700 protein spots were detected with silver staining. After N-terminal microsequencing, 7 proteins were identified, which includes ATP synthase α, δ, γ subunits (96, 97).

*Synechocystis* sp. PCC6803 is a unicellular cyanobacterium. The organism contains large amount of thylakoid membranes, of which many proteins function in photosynthesis or respiration. Thus it is a good model system for proteomic study. The proteomic study of *Synechocystis* was firstly performed by Takashi Sazuka (82, 98). Using sucrose gradient ultracentrifugation, they separate thylakoid membrane and insoluble fraction from soluble fractions. Subsequently, the proteins were extracted with lysis buffer containing 9M urea, 2% NP-40, 2% 2-mercaptoethanol, 0.8% carrier ampholytes, and 2% n-heptyl-β-D-thioglucoside prior to be separated by 2-DE; the resolved protein spots in 2-D gels were then electroblotted onto a PVDF membrane and subjected to N-terminal sequencing. Twenty two proteins were identified as thylakoid membrane proteins and insoluble proteins. Another proteomic study on thylakoid membrane proteins has also been conducted in our group (9). In this report, thylakoid membrane was isolated by differential centrifugation. The total membrane proteins were precipitated with 10% TCA in acetone; lipids were removed by extensive washing with ice-cold acetone. The proteins were then extracted with a multiple chaotropic IEF buffer. After 2-DE, about 200 proteins spots were detected on a 18 cm SDS-PAGE gel stained with colloidal Coomassie blue. The protein spots were excised and subjected in-gel digestion with trypsin. The peptide mass finger printing generated with MALDI-TOF mass spectrometry was used to search database with MS-Fit program, 78 proteins were identified as derived from 51 individual genes. Seventeen of them are hypothetical proteins. Topological analysis shows that most of the proteins are peripheral proteins. In order to resolve more integral proteins on 2-D gel, serial extraction was also performed. EDTA buffer was used to remove cytoplasmic proteins from thylakoid membrane. Peripheral proteins were extracted with 8 M urea prior to be precipitated with 10% TCA in acetone and followed by dissolving with multiple-chaotropic buffer. Integral proteins were washed with ice-cold acetone to remove lipids and pigments followed by extraction with multiple-chaotropic buffer. Both of the proteins were resolved into 2-D gels. In the 2-D gel with pH 3-10, about 400 protein spots were detected with silver-staining. Ninety of them
were identified as proteins derived from 58 individual genes. In the 2-D gel of pH 3-10 for integral proteins, about 100 proteins were detected; thirty of them were identified as derived from 17 genes. By so far, we totally identified 122 single gene derived proteins. Fifty eight of them were overlapped with the proteins in CyanoBase (unpublished data). Topological analysis shows that most of the proteins have one or more transmembrane domain. In order to determine whether acetone-washing can cause the loss of integral proteins, different washing methods were also performed. After peripheral fraction was extracted with 8M urea, the remaining pellet was washed with acetone or water respectively; the pellet was then extracted with 1% SDS at room temperature followed by separation with SDS-PAGE. We found that the gel patterns were almost the same for different processing. Some mark integral protein such as PsaA-B, which always present in large quantity in SDS-PAGE when total membrane were extracted with 1% SDS, were missed on the SDS-PAGE. The data means that some integral membrane proteins were denatured when the membrane were extracted with 8M urea, the hydrophobic core of the integral proteins were exposed, which lead to protein aggregation. The aggregated proteins are refractory to any solubilizing solution, even the most powerful solubilizing solution containing SDS. Acetone may also cause some aggregation, but the loss of integral proteins caused by acetone-washing is trivial compared with the lost caused by denaturation.

**Yeast** No proteomic study conducted specifically on cell membrane of yeast has been reported. The proteome study of *Saccharomyces cerevisiae* conducted using total extract of yeast cells have identified 169 protein spots, Only 5 proteins have been identified as integral membrane proteins, and 7 as peripheral proteins. The numbers are trivial compared with the predicted number of integral and peripheral proteins, which are 645 and 98 respectively (99). The under representation of membrane proteins may due to the low abundance, enrichment procedure may be required to study the membrane proteome of yeast cells.

5.2. Identified membrane proteins from high plants

The membrane system of plant cells contains plasma membrane, organelle envelope membrane and photosynthetic thylakoid membrane. As plant genomes are more complex
than prokaryotic genomes, the genome sequence of any plants has not been unraveled except *Arabidopsis thaliana*, which has just recently been completely sequenced (100-103). Therefore, the proteome study conducted in plants fall far behind those conducted in prokaryotes, even though more and more EST databases are available now.

*Arabidopsis*  *Arabidopsis thaliana* is the only plant with available complete genome sequence. The relatively less complex genome (120 Mb) encodes 25, 498 genes (100), only 10% of these genes have been studied. Proteome study conducted on both plasma membrane and membrane bound organelles have been reported. Plasma membrane has been isolated from *Arabidopsis thaliana*; about 700 protein spots were resolved into 2-D gel after phase partition with Triton X-114 (90). For the 82 picked up and microsequenced protein spots, half of them showed sequence similarity to proteins of known function, and most of the identified proteins are extrinsic proteins. More recent proteome studies on plasma membrane used several methods to optimize the solubilization procedure. Several membrane proteins were identified including a plasma membrane intrinsic protein 1 (PIP1) with 6 TMs, the H\(^+\)-ATPase with 10 TMs and the V-ATPase 16 kD subunit with 4 TMs and other known hydrophobic proteins such as water channels, cellulose and etc.(53, 66). For the study of organelles proteome, *Arabidopsis thaliana* callus culture has been used instead of leaf cells to eliminate the high abundant photosynthetic proteins. Mitochondrial membrane proteins such as ATP synthase delta chain, processing peptidase, non-ER secretory system membrane proteins such as V-ATPase subunit, putative vacuolar sorting receptor homologous and some other proteins such as BP-80/AtElp receptor, cytochrome b5 and calnexin were identified (104).

*Spinach* Chloroplast membrane proteins have been studied, peripheral and integral membrane proteins have been separated with chloroform/methanol extraction. Nine hydrophobic chloroplast envelope membrane proteins and five thylakoid membrane proteins were identified; most of the identified proteins contain more than one TM (92).

Another proteome research for membrane proteins in higher plants has also been reported. From nitrogen fixing root nodules, Soybean peribacteroid membrane was isolated.
After N-terminal sequencing, 17 putative proteins were identified, eight of them have homologous of known proteins, and all of them are peripheral proteins (105).

5.3. Identified membrane proteins from human and animal cell lines.

More and more study on human cell lines proteome has been reported in recent years. For human colon carcinoma cell line (LIM 1215), proteomic research with non-2-DE methods have been performed. The proteins from membrane were extracted and subjected to SDS-PAGE, and then in-gel digested. The proteins were identified with capillary chromatography reversed phase high performance liquid chromatography (RP-HPLC) coupled with electrospray ionization-ion trap mass spectrometry (ESI-IT-MS), 284 proteins were identified, 92 of them are membrane proteins (106). For human monocyte derived U937 cell lines, 2-DE and mass spectrometry were used to study the membrane proteome. Thirty proteins were identified; eight of them are peripheral proteins (107). For human spermatozoa membrane, the proteins were resolved in 2-D gel, and then identified with antibody. Fourteen surface antigens were detected (108).

Other proteomic study conducted on organelles membrane of animal cells has also been reported. For example, the proteome of the rhodopsin-bearing post–Golgi compartment of retinal photoreceptor cells from Southern leopard frogs (109), the proteome of rat mitochondrial membrane (94), and the proteome of rat liver Golgi complex (93). In each case, several hydrophobic membrane proteins have been identified.

6. Genome-wide structure and functional analysis of integral membrane proteins

6.1. Topological prediction for single membrane protein

The structure of membrane proteins contains three parts, i.e., the membrane spanning segments, the cytoplasmic loop and the non-cytoplasmic loop. There are two class of integral membrane proteins based on the type of transmembrane domains (TMs) they have. The TMs of the first class membrane proteins are α-helices, typically with the length of 17 to 25 amino acid residues (110). Most of the membrane proteins found so far belongs to this class. This kind of membrane proteins usually contains one or more membrane spanning α-helices, and the
multiple α-helices sometimes form a bundle in which non-hydrophobic residues may interact with each other between helices. The other class of integral membrane proteins contains β-barrel membrane spanning structure. The only known protein in this class is bacterial outer porins which contain 16 stranded β-barrel structures (111). Little structure prediction has been performed for this class of membrane proteins.

The insolubility in aqueous solutions makes membrane proteins hard to be crystallized. Therefore, the experimental method using X-ray crystallography to determine protein structure is not suitable for integral membrane proteins; even thought they are becoming routines for globular proteins (112). Fortunately, computational prediction of the structure for membrane proteins has been shown to be much easier than for globular proteins and been widely accepted as a routine tool, because the degrees of freedom for the 3D structure of membrane proteins are strongly constrained by the lipid bilayer. Furthermore, the fact that most transmembrane α-helices is formed by unusually long stretch of hydrophobic residues make it easier for structural prediction of membrane proteins (113-119).

The traditional structure prediction methods are based on two major principles. The first one is the hydropathy analysis. As we discussed above, membrane proteins usually contain one or more membrane spanning α-helices with length of 17 to 25 amino acids, the amino acid residues in the α-helices are predominately hydrophobic, the average hydrophobicity of transmembrane α-helices will be much higher than the non-transmembrane part. Based on this fact, a sliding window or a averaged trapezoid window with a size approximately equal to the length of transmembrane α-helices can be used to scan protein sequences, and an empirical cutoff for hydrophobic score will be used to determine which one is the transmembrane α-helix. The second principle used is the so-called “positive inside” rule. It has been found that the amino acid composition of integral membrane proteins is very different between the cytoplasmic side and the non-cytoplasmic side. Positive charged residues such as arginine and lysine residues are mainly found on the cytoplasmic side. This rule can also be applied to the organelle membrane. Based on positive inside rule, the location and orientation of transmembrane α-helices can be accurately predicted. The first algorithm based on these two principles to predict topology of membrane proteins is TopPred.
(117) (available at http://www.expasy.eh); this method applied two empirical hydrophobicity
cutoffs to the output of trapezoid sliding window to distinguish certain transmembrane
helices from putative helices. The best prediction was the helix with the strongest enrichment
of positively charged residues in loops on the cytoplasmic side, loops longer than 70 amino
acid residues were ignored.

Although the algorithms based on the two principles worked well for the structure
prediction of most integral membrane proteins, the accuracy is not as high as we expected.
The reason is that positive inside rule sometimes can be interfered by the globular domains in
loops on the non-cytoplasmic side that contain large number of positively charged residues.
Furthermore, although it has been believed that positively charged residues in short loops
guide the orientation of helices by preventing translocation across the membrane, long loops
containing positively charged residues often fail to prevent the translocation (120, 121).
Besides that, the frequency of Arg and Lys in translocated parts are not reduced significantly
for eukaryotic plasma membrane proteins (122).

Several other prediction algorithms have also been widely adopted. Persson et. al.
developed an algorithm based on the statistics procedure (123). The algorithm took the
advantage of multiple sequence alignment. By finding protein homologs from many sets of
protein family by sequence alignment, the probabilities of each amino acid residues that
appear in the transmembrane helix would be determined. After constructing a chart showing
the amino acid compositions for 15 N- and C- terminal alignment positions relative to the
center of the transmembrane segment, a striking residue distribution pattern was found. From
the center to the position 10 of N-terminus and from the Center to the 11 position of C
terminus, the amino acid residues are predominately hydrophobic, thus the length of
transmembrane segment could be determined as 21. Each amino acid residues in the
transmembrane segment would be assigned two propensity values, one for the middle
hydrophobic potion and the other for the terminal regions of the transmembrane segment.
After that, average propensity values were calculated for each position of transmembrane
segment. The core of the transmembrane segment could be determined with an Eight-residue
segments, the core could be elongated if the middle propensity values are higher than a
threshold. Only the segment of length of 15-29 amino acid residues would be considered a
transmembrane helix. This method is shown to be more accurate than methods based on single sequence (123, 124). The other prediction algorithm based on multiple sequence alignment has also been proven to be successful. The algorithm use a neural network system to predict location of transmembrane helices, and 95% accuracy can be achieved (113). A revised version of this algorithm has been developed to apply a dynamic programming-like algorithm to optimize helices compatible with the neural network output (125).

6.2. Genome-wide topology prediction for membrane proteins

With the development of high throughput DNA sequencing technology, more and more genomic sequence will be available in the near future. To date, about 1000 virus and 112 microbial genomes have been sequenced including the genome of Escherichia coli, Haemophilus influenza, Mycobacterium tuberculosis, Synechocystis sp. Bacillus subtilis, Saccharomyces cerevisiae and etc. More complex high plant and animal genomes such as Arabidopsis and human genome are also available or close to be available. Therefore, it is becoming feasible to predict membrane proteins and their topological structure based on the translated genomic sequence. The pioneer work in this area has recently been conducted by Wallin et al.(5). Genomic-wide data (ORFS) were extracted from eubacterial, archaean, and eukaryotic respectively. TopPred program was used to predict the location and the number of transmembrane helices. The results show that 20-30% of all ORFs encode integral membrane proteins. Interestingly, the prediction found a tendency that proteins of unicellular organisms containing multiple transmembrane helices prefer to have 6 or 12 transmembrane helices while proteins of multi-cellular organism such as Caenorhabditis elegans and Homo sapiens tend to have 7 transmembrane helices. A remarkable correlation between genome size and the fraction of ORFs encoding integral membrane proteins were found. The larger genome contains a larger fraction of ORFs. A reasonable explanation to the correlation is that organisms with larger and complex genome usually have more compartments (organelles), thus more membrane proteins are required to form the membrane that would surround those compartments. The other strong correlation between topology and the distribution of positively but not negatively charged residues were also found. Both eubacterial and Eukaryotic membrane proteins follow the positive inside rule, and the bias for (Arg+Lys) is higher than for either of these two residues taken separately, which may suggest that Arg and
Lys tend to locate at the same side of proteins. For Archaean membrane proteins, *Archaeglobus fulgidus* has a strong bias for (Arg+Lys) while *Methanococcus thermoautotrophicum* has a very weak bias for (Arg+Lys). For all organisms, most of the proteins have a small number of transmembrane helices; very few proteins have more than 12 transmembrane helices. With the increasing of the number of transmembrane helices, the number of proteins falls off rapidly (126,127). The contour plots showing the frequency of proteins with a given number of predicted transmembrane helices and given overall length presents a very interesting result, the number of the proteins with large number of transmembrane helices and short connecting loops and the proteins with only one or two transmembrane helices and large extra-membranous domains are far more than the proteins with multiple transmembrane helices and large extra-membranous domains. This pattern is stronger in eubacteria and archaea than in eukaryotes.

**6.3. Genome-wide functional study of membrane proteins**  
With more and more genomic sequence data available, functional study of predicted proteins at a genome scale is becoming a hot topic in modern biology. Obviously, experimental approaches such as mutagenetic or biochemical analysis are not sufficient because of the time and money costing and labor intensity. It is almost impossible to study the function of hundreds of thousands predicted and more being predicted unknown proteins. Therefore, alternative tools other than experimental approaches began to be explored for functional analysis of hypothetical proteins. A functional study of *E.coli* genome has recently been reported. The study was based on the hypothesis that proteins with sequence or structure similarity may have the same function. So a hypothetical protein with unknown function may be assigned a general function similar to a protein of known function with similar sequence or structure. About 30% of the proteins encoded by *E.coli* genome can be predicted as homologs to proteins with known structure. The other genome-wide prediction of protein function explored much more prediction standards. The method uses not only the sequence and structure similarity, but also functional properties to predict function of unknown proteins. Proteins that participate in common structure complex, such as photosystem I and II complexes, metabolic pathways, biological processes or closely related physiological functions can be associated with each other and be assigned a general function (128). Protein
function can also be predicted from genome-wide protein interaction maps. Protein interaction maps could be constructed not only with traditional experimental methods, such as protein affinity chromatography, affinity blotting, co-immunoprecipitation, chemical crosslinking, yeast two-hybrid system and etc.(129), but also can be constructed from evolution information (130). The basic idea of this method is that some genes may fused together during the long time of evolution process because of selective pressure in some organisms, but in some other organisms, the gene fusion events may not happened. Thus functional domains of one fused gene in one organism may exist as several independent genes in other organism, and these independent genes may acquire normal function by interacting with each other. By this approach, 215 genes or proteins in the complete genomes of Escherichia coli, Haemophilus influenzae and Methanococcus jannaschii are predicted to be involved in 64 unique gene fusion events, and the number could be substantially increased if more genome sequence had been used as control.

No specific genome wide functional study of membrane protein has been reported. The reason may due to the fact all the current prediction approaches predict function of proteins by trying to construct links between a hypothetical protein and its functional homologs with known function. Unfortunately, less membrane proteins have been functionally characterized because of their insolubility and therefore little functional information could be obtained from databases. Thus it is very hard to link a hypothetical membrane protein to a protein with known function. However, the sparse database information should not be an excuse for us to withdraw from functional study of membrane proteins. Actually, we are not in the situation of desperately helplessness. We still can seek help from other aspects. First, we can link unknown proteins with biological membrane using structural prediction; then we can make sure that the function these proteins should be membrane related; even for the proteins have no other function but only exist on the membrane as a structural purpose. Thus we can try to relate the hypothetical membrane proteins to the known membrane functions such as photosynthesis, respiration, transporters, channels, signal transduction, receptors and so on. To achieve this goal, the first step is trying to link the hypothetical proteins to known membrane proteins with the methods as we discussed above. If the method works, then we will have a more clear direction to conduct
further study. If it doesn’t work, we still have other choice. For multi-cellular eukaryotic organisms, it will difficult to conduct further study, but for unicellular organism, it is easy to make knockout mutants. The gene encoding the hypothetical protein will be removed from the genome. We can examine the phenotype related with membrane function, for example, we can measure the biosynthetic function with growth experiment; we can check the variation of photosynthetic activity with oxygen uptake and evolution experiment. For other function such as ion transportation across membrane, signal transduction, corresponding experimental methods could also be adopted to examine the variation between wild type cells and knock out mutants. Furthermore, knock out mutants of integral membrane proteins may have serious effect on the lipid composition of membrane, so we can use GC-MS to examine the variation of lipids in membranes. All of the results obtained from unicellular organism could be used for functional study of multi-cellular organisms based on the computational prediction methods we discussed above.

7. Future work for membrane proteomics

Since the biggest problem of membrane proteomic study encountered is the solubility of membrane proteins, the first important task is to resolve many more membrane proteins in 2-D gels. More powerful solubilizing solutions with new detergent should be explored to extract more hydrophobic membrane proteins into solutions. Current 2-DE system has been proven not optimal for resolving some membrane proteins. Precipitation at the position of pI during IEF, aggregating during IEF, loss of extreme high molecular weight and low molecular weight proteins when transferring from the first to second dimension are still remain to be overcome. In order to solve these problems, we need optimize our IEF solutions by using more powerful detergent and reducing reagents, also higher quality of immobilized pH gradients for IEF are required. The procedure to run IEF may also have adverse effects. Long running time may cause some proteins to precipitate; insufficient running time may cause the horizontal streaking. Furthermore, detection and identification of low abundant membrane proteins still remain to be a big problem. Although higher sensitive staining methods could detect more membrane proteins in 2-D gels, low quantity is still an obstacle to further identification. Narrow pH range IPG can somehow alleviate this problem by enrich
proteins in a narrow pH region, but the quality of commercial available narrow pH range IPG are not trouble free, horizontal streaking is a common phenomena when using narrow immobilized pH gradients. One of the best ideas to study the low abundant membrane proteins is to remove those high abundance proteins, thus we can load much more samples when running IEF without loading beyond the limitation. In order to remove high abundant proteins, we can use antibody to precipitate them and removed them from other proteins. Or we can directly knock out the genes encoding those high abundant proteins, thus the proteins could be totally removed from the proteome. The last and most important task is to study the function of membrane proteins at genome wide scale. Although we already have many approaches to predict function of proteins, but the accuracy still need to be improved. By so far, no specific methods for predicting function of membrane proteins have been developed. Therefore, reliable automated tools for functional analysis of membrane proteins will be a dream for proteome researchers in the near future. This is an exciting issue in scientific field; we will have much challenge and much achievement.

References


CHAPTER 3. PROTEOMIC STUDY OF THE PERIPHERAL PROTEINS FROM THYLAKOID MEMBRANES OF THE CYANOBACTERIUM SYNECHOCYSTIS SP. PCC6803

A paper published in Electrophoresis

Yingchun Wang, Jun Sun and Parag R. Chitnis

Summary

The thylakoid membranes of cyanobacteria and plants contain enzymes that function in diverse metabolic reactions. Many of these enzymes and regulatory proteins are associated with the membranes as peripheral proteins. To identify these proteins, we separated and identified the peripheral proteins of thylakoid membranes of the cyanobacterium Synechocystis sp. PCC6803. Trichloroacetic acid (TCA)–acetone extraction was used to enrich samples with peripheral proteins and to remove integral membrane proteins. The proteins were separated by two dimensional-gel electrophoresis (2-DE) and identified by peptide mass fingerprinting. More than 200 proteins were detected on the SDS-PAGE gel that was stained with colloidal Coomassie blue. We analyzed 116 spots by peptide mass fingerprinting and identified 78 spots that were derived from 51 genes. Some proteins were found in multiple spots, indicating differential modifications resulting in charge differences. Therefore, a significant fraction of the peripheral proteins in thylakoid membranes is modified post-translationally. In our analysis, products of 17 hypothetical genes could be identified in the peripheral protein fraction. Therefore, proteomic analysis is a powerful tool to identify location of the products of hypothetical genes and to characterize complexity in gene expression due to post-translational modifications.

Keywords: Cyanobacteria / Proteomics / Peptide mass fingerprinting / Thylakoid membranes / Two-dimensional electrophoresis

1 This work is supported in part by grants from Iowa Corn Promotion Board and from the Vice-Provost for Research at ISU (Promag and SPRIG). Journal Paper No. J-18936 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3416; supported in part by Hatch Act and State of Iowa funds.

2 Reprinted with permission of Electrophoresis, 2000, 21, 1746-1754.

3 Correspondence: Dr. Parag R. Chitnis, 4156 Molecular Biology Building, Iowa State University, Ames, IA 50011, USA; E-mail: chitnis@iastate.edu; Fax: +1-515-294-0453
1 Introduction

Thylakoid membranes of cyanobacteria and chloroplasts contain machinery for oxygenic photosynthesis, which is the major source of biological energy and oxygen on this planet. The pigment-protein complexes of oxygenic photosynthesis harvest light and use its energy for electron transport, generating ATP and NADPH, which are used in many physiological processes. Besides photosynthetic proteins, thylakoid membranes contain a plethora of proteins that are involved in membrane biogenesis, crucial biosynthetic pathways, membrane transport and regulation of energy metabolism. Thus, the proteome of thylakoid membranes plays critical roles in photosynthetic metabolism.

Cyanobacteria have served as a model system for studying metabolic activities related to chloroplasts. Cyanobacteria and chloroplasts have a common ancestor in evolution. Therefore, the metabolic functions in cyanobacteria are similar to those in the chloroplast [1]. The mesophilic unicellular cyanobacterium *Synechocystis* sp. PCC 6803 has been particularly useful in many recent studies. The genome of *Synechocystis* is small (3.57 Mbp) compared to those of high plants and has been sequenced completely [2]. The *Synechocystis* genome contains 3168 genes. The proteome of *Synechocystis* has been studied by 2D electrophoresis and N-terminal amino acid sequencing. This approach has identified 227 protein spots that are derived from 143 genes [3]. Only 13 of these proteins are localized in the thylakoid membranes. A large proportion of genes in the genome of *Synechocystis* sp. PCC 6803 are considered hypothetical genes because they code for putative proteins with unknown functions. To identify the hypothetical genes that are expressed and targeted to proper subcellular compartment, a combination of cell fractionation, proteomics, and reverse genetics provides a powerful tool. We have embarked on a systematic analysis of the proteome of thylakoid membranes.

Thylakoid membranes contain peripheral and integral membrane proteins. Because of the difficulties in separation of integral membrane proteins by 2-DE [4,5], we initially optimized a rapid method for isolation and display of the peripheral proteins. Most of the peripheral proteins are expected to perform many crucial enzymatic and regulatory functions in photosynthetic metabolism. Here we describe 2D electrophoretic separation of proteins
that are associated with thylakoid membranes and their identification by peptide mass fingerprinting.

2 Material and Methods

2.1. Growth of Synechocystis sp. PCC 6803 and preparation of membranes

The wild type strain of Synechocystis sp. PCC 6803 was grown in BG11 medium at 30°C, and cells were harvested at a late exponential phase. The cells were pelleted and resuspended in SMN buffer 0.4mM sucrose, 10mM NaCl, 50 mM MOPS, pH7.0, 5mM EDTA and 0.5mM PMSF. Cells were broken using a bead beater and thylakoid membranes were isolated by differential centrifugation [6]. These preparations are not contaminated significantly by cell membranes and are used routinely in photosynthetic studies. Chlorophyll concentration of the thylakoid preparation was measured with 80% acetone [7].

2.2. Isolation of peripheral proteins from thylakoid membranes

Total proteins in a membrane sample were precipitated with 10% Trichloroacetic acid and 0.1% 2-mercaptoethanol in acetone, and then extracted with 100% acetone for three times to remove pigments and lipids. The protein pellet was dried under vacuum, resuspended in the multiple chaotrope rehydration buffer [8] and sonicated to extract proteins. The proteins solubilized in the rehydration buffer were collected by centrifugation in a microfuge for 15 minutes. This fraction predominantly contains the peripheral membrane proteins. The concentration of the proteins was measured with Bio-RAD Dc protein assay kit (Bio-RAD).

2.3. 2-DE of peripheral proteins from thylakoid membranes

The immobilized pH gradient strips (18cm, pH 3-10, nonlinear gradient, Pharmacia Biotech) were equilibrated with 500μl sample containing approximately 500 μg proteins for 16 h. The first dimension electrophoresis was performed with a multiphor instrument with voltage settings of 300V for 1 h, 1000V for 0.5 h, 2500V for 0.5 h, and 5000V for 16h (total of 82000 V-h). Upon electrophoresis, the proteins on the strips were denatured and cysteinyl residues were reduced [8]. The second dimension electrophoresis was performed on a 12-18% gradient gel. Upon electrophoresis, the protein spots were stained with colloidal
Coomassie blue and the gels were scanned to obtain images for analysis with Melanie II software [9].

2.4. Mass spectrometry

Protein spots from the stained gels were excised and transferred to a 96 well plate using a robot [10]. The excised gel fragments were incubated with 50% acetonitrile to remove dye that was bound to the protein. The gel fragments were dried under vacuum and then incubated for 16 h at 37 °C with 10 μl of 10 μg/ml Trypsin in 2.5 mM Tris-HCl, pH 8.5. The resulting tryptic fragments were eluted in 50% acetonitrile and 0.5% trifluoroacetic acid by diffusion that was facilitated by ultrasonication. The tryptic peptide fragments were used for MALDI-TOF mass spectrometry using 2 μl of each sample, alpha-cyano-4-hydroxycinnamic acid as a matrix, and Micromass TofSpec 2E mass spectrometer (Micromass, Manchester UK). Spectra were obtained in reflectron delayed extraction mode over a mass range of 800 to 4000 Da. Spectra from 64 shots at 20 different positions were combined to generate a peptide mass fingerprint for each protein sample. The peptide ions generated by autolysis of trypsin (with m/z 842.51+ and 2211.10+) were used as internal standards for calibrating the mass spectra.

2.5 Data Analysis

Peptide masses of peptide were analyzed with MS-FIT program (http://prospector.ucsf.edu) using the following parameters: mass tolerance of 0.1Da, a minimum of four peptide matches, and no missed cleavages. When isoforms were found for some proteins, FindMod and PeptideMass tools in ExPaSy were used to predict possible protein modifications [11]. Hydropathy analysis of the deduced sequences of the novel proteins was performed using TopPred2 program [12], whereas the homology search was performed using BLAST program.

3. Results

3.1. Sample preparation

Integral membrane proteins are underrepresented in the current proteomic databases. To investigate causes of the loss of membrane proteins during sample preparation, we
compared commonly used methods of sample preparation and examined the protein yield by SDS-PAGE. Figure 1 shows a silver-stained SDS-polyacrylamide gel with protein samples after different treatments. As a control, we denatured membranes directly (Fig. 1, first lane on left). Many integral membrane proteins and peripheral proteins can be identified from the apparent migration. For example, PsaA-PsaB are the polytopic core proteins of photosystem I and form a major band at ~68 kDa [13]. Similarly, CP47 is a 47-kDa integral membrane protein of photosystem II (PSII) [14]. In contrast, other two major bands of 18 and 8 kDa contain PsaD and PsaC, respectively [13]. To examine the effects of protein extraction without any additional treatments, membranes were extracted at room temperature with rehydration buffer and the proteins were separated by PAGE (Fig. 1). Compared to the control lane, the high concentration of Triton X-100 in rehydration buffer affected migration of many proteins in the sample. However, both samples contained similar levels of proteins. Therefore, rehydration buffer can solubilize the integral membrane proteins. When we used solubilization with SDS prior to extraction with the rehydration buffer, extraction and mobility of membrane proteins were not improved (Fig. 1).

Precipitation with TCA is a commonly used step in sample preparation for 2-DE. To test the effects of TCA precipitation on extraction, membranes were incubated at room temperature with rehydration buffer or water and then extracted with TCA. Figure 1 shows that extraction of membrane proteins followed by TCA precipitation caused a severe loss of membrane proteins. Therefore, when denatured proteins are precipitated, they formed insoluble aggregates. In contrast, when membranes were subjected to TCA precipitation directly, a selective fraction of proteins could be recovered quantitatively. Examination of the protein pattern showed that the peripheral proteins, such as PsaD and PsaC, could be solubilized after TCA precipitation whereas the precipitated integral membrane proteins could not be dissolved in rehydration buffer. Treatment of pellets with SDS before extraction with rehydration buffer did not improve the recovery of integral membrane proteins. Therefore, TCA precipitation provides an effective and rapid method for removing most of the integral membrane proteins from the thylakoid sample. Urea and detergent treatment increased the amount of proteins in samples, but the samples contained pigments and other lipid-soluble components, which could have a detrimental effect on protein migration,
resolution and visualization. Use of sequential extraction of protein samples [8] did not provide any significant advantage for thylakoid membrane proteins (data not shown). Therefore, we used TCA precipitation followed by extraction with multiple chaotrope buffer for routine 2D analysis of the peripheral proteins of thylakoid membranes.

3.2. 2-D separation of proteins

The soluble proteins were separated according to their pI using IEF and further resolved according to their mass with SDS-PAGE. When we used a linear pH gradient (pH 3-10), we found that most of the proteins were resolved in the pH 4-7 range. Fewer protein spots were observed in the high pH range than in the pH 4-7 range. When we used strips with a nonlinear pH gradient, which has a less steep gradient in the pH 4-7 range, better separation of the proteins with a pI between 4-7 resulted. Therefore, we selected pH 3-10 nonlinear gradient strips for subsequent experiments.

To analyze the protein spots on the 2-D gel, we used the Melanie II program to detect protein features in gel images. Known proteins were used to calibrate apparent masses of proteins in a gel. The pI values were deduced from the linear gradient gels. About 200 proteins can be detected on the 18 cm nonlinear pH gradient 2-D gel, 116 of them were subjected to in-gel digestion with trypsin. Spots from two independent gels were used for mass spectrometry studies. The identity of proteins with high scores in the MS-FIT analysis was further validated with three other criteria: mass, pI and correspondence to the spots in Cyano2Dbase. The deduced mass of the putatively identified proteins should match the protein mass estimated from the migration on the 2-D gel. In the mass comparisons, we considered possible post-translational processing which could increase or decrease the mass. The second criterion was the predicted pI of proteins, which should be close to the pI estimated from the 2-D gel. Again, we considered the possibility of post-translational modifications that may alter pI. Many proteins have isoforms with differences in pI due to post-translational modifications. For example, the protein spots on the 2D gel 85, 86, 88 and 89 differ in pI, but were proven to be isoforms of the PsaD subunit of photosystem I. The last criterion is the correspondence to the identified proteins in the Cyano2Dbase. The proteins in this database have been identified by determining N-terminal amino acid sequences [3,15]. If
a protein on our gel was identified as the same protein on the Cyano2Dbase and was located at the same place on the gel, we considered the identification as correct. For example, the PsaD spots identified on our gel were located at the same place on the 2-D gel of Cyano2Dbase. Our analysis also considered the presence of known proteins and their identity based on Western blotting and N-terminal sequencing. Such proteins have been included in Table 1 despite a low score in analysis of peptide mass fingerprints.

With these methods and criteria, 78 proteins were identified (Table 1). Most of the identified proteins have 10 kDa or higher molecular mass. It was difficult to identify low molecular weight proteins with MS-FIT program, unless the mass tolerance and number of matching peptides were lowered. This may be due to modification of small proteins could give less than four unmodified peptides that were used in identification from peptide mass fingerprints. High molecular weight proteins are expected to contain sufficient number of unmodified peptides with a mass between 800 and 4000Da. Therefore, many peptides generated by trypsin could be fitted with MS-Fit program. However, low molecular weight proteins may yield only a few unmodified tryptic peptides in the mass range used in the analyses here. Some proteins identified here have also been identified by N-terminal sequencing in the Cyano2Dbase (Table 1). However, many proteins identified in this study do not appear in the Cyano2Dbase. Thus, more proteins could be identified by peptide mass fingerprinting than with N-terminal amino acid sequencing, indicating a higher sensitivity of this technique. Despite the higher number of proteins identified in this study, we could identify only 67% spots that were subjected to peptide mass fingerprinting. The remaining spots could not be identified because of the poor quality of samples, less amount of proteins, less number of unmodified peptides in these proteins, or errors in calibration.

Among the proteins identified in our analysis, 11 proteins form 38 spots with different pI values (Table 2). The frequent occurrence of isoforms is consistent with the observations reported for Cyano2Dbase. Post-translational modification of these proteins could have led to different charges on the amino acid residues. We used the FindMod program to search possible modifications on the amino acid residues. Such analyses provide a rapid method for prediction of the nature and site of post-translational modifications. Many isoforms were predicted to contain methylation of basic residues which could lead to
differences in the pI values. Methylation is one of the most commonly found protein modification in a cell [16].

3.3. Products of hypothetical genes

Among 78 identified proteins from the mass fingerprints, 17 were products of hypothetical genes. From the genome sequence of *Synechocystis* sp. PCC6803, the function and intracellular location of these gene-products is not apparent. It is not known if these genes are expressed at all. The proteome analysis identified these gene products in the thylakoid membranes. Therefore, we now know that these genes are expressed and their products associate with the thylakoid membranes. To understand the transmembrane topology of these proteins we performed hydropathy analysis of the deduced amino acid sequences using TopPred2 program. These analyses showed that ten of the products of hypothetical genes do not contain hydrophobic regions that could cross the lipid bilayer. Therefore these are soluble proteins that associate with thylakoid membranes peripherally. Six products of hypothetical genes contain one putative transmembrane helix (Table 3). Interestingly, these hydrophobic regions are located near the N-terminus, which is a characteristic of a presequence for targeting a protein to thylakoid lumen [17]. Thus these proteins could be peripheral proteins in the thylakoid lumen. In contrast, the product of the sll0503 open reading frame may have two transmembrane helices that are located well inside the primary structure of the protein. These hydrophobic regions may cross the membranes. However, our analysis did not detect any of the abundant integral membrane proteins of the thylakoid membranes. Similarly, the products of the genes with identified functions and homology are known to be peripheral proteins or soluble proteins. Therefore, it is likely that the hydrophobic regions that were considered as putative transmembrane helices may indeed be used for interaction with the hydrophobic lipid bilayer rather than for crossing the membrane.

4 Discussion

Membrane proteins constitute about 30-40% of gene products that are encoded by bacterial genomes [18]. Yet, the proteomic databases contain very few integral membrane proteins. For example, Cyano2Dbase contains only 9.1% integral membrane proteins. This
study identifies one problem that leads to the loss of integral membrane proteins during 2D electrophoresis. In the past, many attempts have been made towards better recovery of integral membrane proteins [18-20]. The recovery of integral membrane proteins in these studies was an improvement over the previous methods, but was poor compared to that for soluble proteins. These studies involved TCA precipitation as one of the steps in sample preparation. Our studies demonstrate that TCA precipitation results in the loss of integral membrane proteins. Therefore, we used TCA precipitation for preparation of soluble proteins enriched with peripheral membrane proteins. Of 78 spots identified in these studies, none of the known proteins is an integral membrane protein and only seven of the products of hypothetical genes have potential transmembrane helices. Therefore, alternatives to TCA precipitation are needed to resolve membrane proteins by 2-DE.

Table 4 lists functions of the peripheral proteins that were identified in this study. As expected, many proteins are components of the photosynthetic machinery. Some soluble proteins that are predominantly present in the cytoplasmic fraction were detected in the peripheral protein fraction. For most such proteins, it is unlikely that their presence represents contamination of thylakoid membranes with cytoplasm. We believe that the presence of these proteins in thylakoid membranes represents functional significance. There are several arguments to support our postulate. First, many of these proteins were present in substantial amounts in the thylakoid membrane fraction. Second, proper care was taken during membrane preparation to remove cytoplasm from membranes. We used multiple centrifugation-resuspension steps to remove soluble protein that do not associate with thylakoid membranes. Third, only selected soluble proteins were found in the peripheral protein fraction. With the exception of some like Rubisco small subunit, the abundant soluble proteins were absent in the peripheral membrane protein fraction. In addition, we did not find any of the abundant soluble proteins, such as DnaK and GroEL in the thylakoid membrane proteins. For many soluble proteins, their presence in the membrane fraction represented the tendency of these proteins to associate with membranes to carry out their function. For example, phycobilisomes are known to associate with the photosystem II complexes in the thylakoid membranes [1]. Therefore, the identification of allophycocyanin and phycocyanin in the peripheral proteins was expected [21]. Many proteins involved in gene expression (e.g.
ribosomal proteins) were also present in the peripheral membrane fraction. Association of ribosomes and transcription machinery with the thylakoid membranes of chloroplasts has been reported [22]. Our results suggest that polysomes may also associate with the cyanobacterial photosynthetic membranes and assist in protein integration into membranes.

Our research shows that mass finger printing is a useful tool to identify proteins on a large scale. We detected 200 protein spots in the gel and identified 78 protein spots in the 2-D gel. A major finding of the present study is the detection of many products of hypothetical genes in thylakoid membranes. Many of these newly identified proteins of thylakoid membranes showed substantial regions of similarity to the products of plant ESTs (data not shown). As an example, we discuss here the homology between sll2501 and AAC13631 of Arabidopsis. The hypothetical gene sll2501 of Synechocystis codes for a small hydrophilic protein with 89 amino acid residues. The Arabidopsis cDNA codes for a protein of 116 residues. A region of 81 amino acids contains 28.4% identical residues and additional 25.9% conservative replacements. It is likely that the N-terminal extension in the Arabidopsis protein targets the protein to chloroplasts, where it may associate with the thylakoid membranes. To demonstrate that the other ESTs code for homologous proteins may require complete sequences of these cDNAs. Targeted inactivation of the hypothetical genes identified in this study is in progress and will allow us to identify role of these proteins in biogenesis and function of thylakoid membranes. To summarize, proteomic analysis presented here identified TCA precipitation as a problem in the analysis of membrane proteins, demonstrated common occurrence of post-translational modifications in thylakoid proteins, and revealed many new proteins that are components of the thylakoid membranes.

We thank Wu Xu and Larysa Porubleva for assistance with 2-DE and Tina Haupt for data search. PRC thanks Dr. Brad Walsh and his wonderful associates at the Australian Proteome Analysis Facility for their hospitality, use of the facilities, and superb technical training in proteomics. This work is supported in part by grants from Iowa Corn Promotion Board and from the Vice-Provost for Research at ISU (Promag and SPRIG). Journal Paper No. J-18936 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3416; supported in part by Hatch Act and State of Iowa funds.
5. Reference

[1] Bryant, D.A. in: Platt, T. and Li, W.K.W., (Eds.), The cyanobacterial photosynthetic apparatus: comparison to those of higher plants and photosynthetic bacteria, Department of Fisheries and Oceans, Ottawa, Canada 1986, pp. 423-500


<table>
<thead>
<tr>
<th>No.</th>
<th>Protein ID</th>
<th>Deduced Mass</th>
<th>Deduced pl</th>
<th>apparent pl</th>
<th>Peptides Matched</th>
<th>Aas covered by the matched peptides</th>
<th>Mascot score</th>
<th>NcBI accession No.</th>
<th>Cyanobase No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chloroplast import-associated channel IAP75 protein D</td>
<td>62307.8</td>
<td>4.29</td>
<td>4.08</td>
<td>15</td>
<td>22%</td>
<td>1.28E+08</td>
<td>1652591</td>
<td>slr1227</td>
</tr>
<tr>
<td>2</td>
<td>ATP synthase b subunit</td>
<td>57735.4</td>
<td>4.81</td>
<td>4.4</td>
<td>13</td>
<td>49%</td>
<td>1.05E+08</td>
<td>1653357</td>
<td>slr1270</td>
</tr>
<tr>
<td>3</td>
<td>ATP synthase b subunit</td>
<td>51733.3</td>
<td>4.89</td>
<td>4.89</td>
<td>16</td>
<td>49%</td>
<td>1.60E+11</td>
<td>114570</td>
<td>slr1329</td>
</tr>
<tr>
<td>4</td>
<td>ATP synthase b subunit</td>
<td>51733.3</td>
<td>4.89</td>
<td>4.98</td>
<td>25</td>
<td>72%</td>
<td>3.46E+14</td>
<td>114570</td>
<td>slr1329</td>
</tr>
<tr>
<td>5</td>
<td>ATP synthase b subunit</td>
<td>51733.3</td>
<td>4.89</td>
<td>5.01</td>
<td>17</td>
<td>53%</td>
<td>1.12E+12</td>
<td>114570</td>
<td>slr1329</td>
</tr>
<tr>
<td>6</td>
<td>ATP synthase b subunit</td>
<td>51733.3</td>
<td>4.89</td>
<td>5.16</td>
<td>14</td>
<td>50%</td>
<td>1.25E+10</td>
<td>114570</td>
<td>slr1329</td>
</tr>
<tr>
<td>7</td>
<td>ATP synthase b subunit</td>
<td>51733.3</td>
<td>4.89</td>
<td>5.25</td>
<td>5</td>
<td>21%</td>
<td>5.47E+03</td>
<td>114570</td>
<td>slr1329</td>
</tr>
<tr>
<td>8</td>
<td>Phytoene desaturase</td>
<td>52921.1</td>
<td>5.13</td>
<td>5.3</td>
<td>5</td>
<td>14%</td>
<td>353</td>
<td>1348539</td>
<td>slr1254</td>
</tr>
<tr>
<td>9</td>
<td>ATP synthase b subunit</td>
<td>53965.8</td>
<td>4.96</td>
<td>5.15</td>
<td>5</td>
<td>10%</td>
<td>2.18E+03</td>
<td>114530</td>
<td>slr1326</td>
</tr>
<tr>
<td>10</td>
<td>ATP synthase b subunit</td>
<td>53965.8</td>
<td>4.96</td>
<td>5.12</td>
<td>8</td>
<td>17%</td>
<td>7.36E+04</td>
<td>114530</td>
<td>slr1326</td>
</tr>
<tr>
<td>11</td>
<td>ATP synthase b subunit</td>
<td>53965.8</td>
<td>4.96</td>
<td>5.25</td>
<td>5</td>
<td>16%</td>
<td>768</td>
<td>114530</td>
<td>slr1326</td>
</tr>
<tr>
<td>12</td>
<td>ATP synthase a subunit</td>
<td>53965.8</td>
<td>4.96</td>
<td>6.34</td>
<td>5</td>
<td>15%</td>
<td>768</td>
<td>114530</td>
<td>slr1326</td>
</tr>
<tr>
<td>13</td>
<td>ATP synthase a subunit</td>
<td>53965.8</td>
<td>4.96</td>
<td>5.46</td>
<td>6</td>
<td>17%</td>
<td>2.32E+03</td>
<td>114530</td>
<td>slr1326</td>
</tr>
<tr>
<td>14</td>
<td>Hypothetical protein</td>
<td>51159.2</td>
<td>4.99</td>
<td>5.52</td>
<td>4</td>
<td>13%</td>
<td>249</td>
<td>1001345</td>
<td>slr0053</td>
</tr>
<tr>
<td>15</td>
<td>Hypothetical protein</td>
<td>36570.2</td>
<td>4.57</td>
<td>4.35</td>
<td>10</td>
<td>40%</td>
<td>5.29E+05</td>
<td>2500385</td>
<td>slr1356</td>
</tr>
<tr>
<td>16</td>
<td>Heat shock protein GspE</td>
<td>27567.7</td>
<td>4.62</td>
<td>4.4</td>
<td>6</td>
<td>29%</td>
<td>901</td>
<td>2459592</td>
<td>slr0057</td>
</tr>
<tr>
<td>17</td>
<td>Hypothetical protein</td>
<td>34015.2</td>
<td>4.95</td>
<td>4.98</td>
<td>8</td>
<td>32%</td>
<td>8.98E+04</td>
<td>1853639</td>
<td>slr0151</td>
</tr>
<tr>
<td>18</td>
<td>Hypothetical protein</td>
<td>29911.8</td>
<td>4.82</td>
<td>4.44</td>
<td>4</td>
<td>25%</td>
<td>1.43E+03</td>
<td>131837</td>
<td>slr0427</td>
</tr>
<tr>
<td>19</td>
<td>Hypothetical protein</td>
<td>27011.9</td>
<td>7.93</td>
<td>4.46</td>
<td>8</td>
<td>38%</td>
<td>5.36E+04</td>
<td>1001106</td>
<td>slr0431</td>
</tr>
<tr>
<td>20</td>
<td>Hypothetical protein</td>
<td>28419.6</td>
<td>4.76</td>
<td>4.62</td>
<td>4</td>
<td>31%</td>
<td>247</td>
<td>3024873</td>
<td>slr0075</td>
</tr>
<tr>
<td>21</td>
<td>Plastoquinol-plastoxygenin reductase</td>
<td>30633.4</td>
<td>5.97</td>
<td>5.4</td>
<td>4</td>
<td>18%</td>
<td>1.50E+03</td>
<td>1853881</td>
<td>slr1868</td>
</tr>
<tr>
<td>22</td>
<td>Plastoquinol-plastoxygenin reductase</td>
<td>18996.5</td>
<td>4.87</td>
<td>4.5</td>
<td>6</td>
<td>50%</td>
<td>4.56E+03</td>
<td>1717958</td>
<td>slr1316</td>
</tr>
<tr>
<td>23</td>
<td>Phytochrome b subunit</td>
<td>18126.6</td>
<td>4.98</td>
<td>4.62</td>
<td>6</td>
<td>50%</td>
<td>2.41E+04</td>
<td>2493300</td>
<td>slr1577</td>
</tr>
<tr>
<td>24</td>
<td>12 KD extrinsic protein of PSII</td>
<td>14231.3</td>
<td>4.44</td>
<td>3.63</td>
<td>5</td>
<td>53%</td>
<td>1.74E+03</td>
<td>1255692</td>
<td>slr1194</td>
</tr>
<tr>
<td>25</td>
<td>12 KD extrinsic protein of PSII</td>
<td>14231.3</td>
<td>4.44</td>
<td>4.26</td>
<td>4</td>
<td>49%</td>
<td>1.74E+03</td>
<td>1255692</td>
<td>slr1194</td>
</tr>
<tr>
<td>26</td>
<td>ribosomal protein L12</td>
<td>13259.4</td>
<td>4.74</td>
<td>4.45</td>
<td>7</td>
<td>65%</td>
<td>1.39E+03</td>
<td>132564</td>
<td>slr1746</td>
</tr>
<tr>
<td>27</td>
<td>PSII 13 KD protein</td>
<td>12590.4</td>
<td>4.96</td>
<td>4.8</td>
<td>4</td>
<td>48%</td>
<td>901</td>
<td>3334306</td>
<td>slr1398</td>
</tr>
<tr>
<td>28</td>
<td>ATPase subunit epsilon</td>
<td>14580.6</td>
<td>5.2</td>
<td>5.61</td>
<td>8</td>
<td>50%</td>
<td>1.76E+05</td>
<td>114608</td>
<td>slr1350</td>
</tr>
<tr>
<td>29</td>
<td>ATPase subunit epsilon</td>
<td>14580.6</td>
<td>5.2</td>
<td>4.89</td>
<td>6</td>
<td>50%</td>
<td>7.51E+03</td>
<td>114608</td>
<td>slr1330</td>
</tr>
<tr>
<td>30</td>
<td>Hypothetical Protein</td>
<td>9170.6</td>
<td>5.23</td>
<td>4.62</td>
<td>4</td>
<td>35%</td>
<td>71</td>
<td>1653051</td>
<td>srr3532</td>
</tr>
<tr>
<td>31</td>
<td>carbon dioxide concentrating mechanism protein ComK</td>
<td>11134.8</td>
<td>5.33</td>
<td>5.63</td>
<td>3</td>
<td>40%</td>
<td>1.38E+03</td>
<td>2493553</td>
<td>slr1028</td>
</tr>
<tr>
<td>32</td>
<td>PII protein</td>
<td>12397.4</td>
<td>6.33</td>
<td>5.7</td>
<td>6</td>
<td>48%</td>
<td>5.88E+03</td>
<td>2494772</td>
<td>slr0707</td>
</tr>
<tr>
<td>33</td>
<td>PII protein</td>
<td>12397.4</td>
<td>6.33</td>
<td>5.97</td>
<td>6</td>
<td>48%</td>
<td>5.88E+03</td>
<td>2494772</td>
<td>slr0707</td>
</tr>
<tr>
<td>34</td>
<td>PII protein</td>
<td>6577.5</td>
<td>5.13</td>
<td>5.16</td>
<td>5</td>
<td>65%</td>
<td>1.82E+03</td>
<td>1673344</td>
<td>slr0382</td>
</tr>
<tr>
<td>35</td>
<td>IMP dehydrogenase</td>
<td>17270.1</td>
<td>5.18</td>
<td>5.34</td>
<td>7</td>
<td>45%</td>
<td>4.81E+05</td>
<td>1001427</td>
<td>slr0172</td>
</tr>
<tr>
<td>36</td>
<td>Hypothetical protein</td>
<td>10946.7</td>
<td>4.63</td>
<td>4.16</td>
<td>6</td>
<td>55%</td>
<td>8.11E+03</td>
<td>1651747</td>
<td>slr0729</td>
</tr>
<tr>
<td>37</td>
<td>Hypothetical protein</td>
<td>10684.1</td>
<td>4.61</td>
<td>4.08</td>
<td>8</td>
<td>64%</td>
<td>2.09E+05</td>
<td>1852665</td>
<td>slr2501</td>
</tr>
<tr>
<td>38</td>
<td>Ribulose bisphosphate carboxylase small subunit</td>
<td>13239.1</td>
<td>5.36</td>
<td>5.16</td>
<td>4</td>
<td>38%</td>
<td>769</td>
<td>1710044</td>
<td>slr0012</td>
</tr>
<tr>
<td>39</td>
<td>PII protein</td>
<td>12397.4</td>
<td>6.33</td>
<td>5.06</td>
<td>5</td>
<td>47%</td>
<td>4.50E+03</td>
<td>2494772</td>
<td>slr0707</td>
</tr>
<tr>
<td>40</td>
<td>PII protein</td>
<td>8288.6</td>
<td>6.56</td>
<td>4.8</td>
<td>4</td>
<td>68%</td>
<td>629</td>
<td>1673312</td>
<td>slr1690</td>
</tr>
<tr>
<td>41</td>
<td>phycocyanin a subunit</td>
<td>17586.7</td>
<td>5.35</td>
<td>4.85</td>
<td>7</td>
<td>59%</td>
<td>4.44E+04</td>
<td>2493297</td>
<td>slr1578</td>
</tr>
</tbody>
</table>
Table 1. (continued)

<table>
<thead>
<tr>
<th>No.</th>
<th>Protein ID</th>
<th>Deduced Mass</th>
<th>Deduced pl</th>
<th>Peptides Matched</th>
<th>Peptides covered by the matched peptides</th>
<th>Mowse score</th>
<th>NCBI accession No.</th>
<th>Cyanobase No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>Hypothetical protein</td>
<td>38270.8</td>
<td>7.82</td>
<td>6.78</td>
<td>12</td>
<td>47%</td>
<td>1.06E+07</td>
<td>sll1306</td>
</tr>
<tr>
<td>72</td>
<td>Hypothetical protein</td>
<td>38270.8</td>
<td>7.82</td>
<td>7.14</td>
<td>13</td>
<td>47%</td>
<td>2.93E+07</td>
<td>sll1306</td>
</tr>
<tr>
<td>73</td>
<td>Hypothetical protein</td>
<td>38270.8</td>
<td>7.82</td>
<td>7.0</td>
<td>9</td>
<td>34%</td>
<td>1.72E+05</td>
<td>sll1306</td>
</tr>
<tr>
<td>74</td>
<td>Heme oxygenase</td>
<td>27050.8</td>
<td>6.24</td>
<td>6.33</td>
<td>8</td>
<td>39%</td>
<td>4.71E+04</td>
<td>sll1184</td>
</tr>
<tr>
<td>75</td>
<td>Light repressed protein</td>
<td>21895</td>
<td>6.07</td>
<td>6.78</td>
<td>8</td>
<td>53%</td>
<td>4.22E+04</td>
<td>sll0947</td>
</tr>
<tr>
<td>76</td>
<td>Light repressed protein</td>
<td>21895</td>
<td>6.07</td>
<td>6.96</td>
<td>9</td>
<td>57%</td>
<td>1.04E+05</td>
<td>sll0947</td>
</tr>
<tr>
<td>77</td>
<td>Light repressed protein</td>
<td>21895</td>
<td>6.07</td>
<td>6.82</td>
<td>7</td>
<td>48%</td>
<td>1.72E+04</td>
<td>sll0947</td>
</tr>
<tr>
<td>78</td>
<td>Global nitrogen regulator</td>
<td>25046.3</td>
<td>6.53</td>
<td>7.2</td>
<td>6</td>
<td>38%</td>
<td>6.72E+03</td>
<td>sll1423</td>
</tr>
<tr>
<td>79</td>
<td>SOS function regulatory protein</td>
<td>22744.3</td>
<td>5.84</td>
<td>6.33</td>
<td>6</td>
<td>39%</td>
<td>1.28E+04</td>
<td>sll1626</td>
</tr>
<tr>
<td>80</td>
<td>Light repressed protein</td>
<td>21895</td>
<td>6.07</td>
<td>6.33</td>
<td>7</td>
<td>48%</td>
<td>1.72E+04</td>
<td>sll0947</td>
</tr>
<tr>
<td>81</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>9.6</td>
<td>4</td>
<td>30%</td>
<td>1.19E+03</td>
<td>sll0707</td>
</tr>
<tr>
<td>82</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>9.5</td>
<td>3</td>
<td>22%</td>
<td>1.19E+03</td>
<td>sll0707</td>
</tr>
<tr>
<td>83</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>9</td>
<td>2</td>
<td>10%</td>
<td>280</td>
<td>sll0707</td>
</tr>
<tr>
<td>85</td>
<td>Light repressed protein</td>
<td>21895</td>
<td>6.07</td>
<td>6.82</td>
<td>7</td>
<td>48%</td>
<td>1.72E+04</td>
<td>sll0947</td>
</tr>
<tr>
<td>86</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>87</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>88</td>
<td>Global nitrogen regulator</td>
<td>25046.3</td>
<td>6.53</td>
<td>7.2</td>
<td>6</td>
<td>38%</td>
<td>6.72E+03</td>
<td>sll1423</td>
</tr>
<tr>
<td>89</td>
<td>SOS function regulatory protein</td>
<td>22744.3</td>
<td>5.84</td>
<td>6.33</td>
<td>6</td>
<td>39%</td>
<td>1.28E+04</td>
<td>sll1626</td>
</tr>
<tr>
<td>91</td>
<td>PII protein</td>
<td>12397.4</td>
<td>6.33</td>
<td>6.96</td>
<td>6</td>
<td>58%</td>
<td>3.97E+03</td>
<td>sll0707</td>
</tr>
<tr>
<td>92</td>
<td>Photosystem I subunit IV linker protein</td>
<td>8145.2</td>
<td>8.03</td>
<td>8.82</td>
<td>4</td>
<td>78%</td>
<td>2.09E+02</td>
<td>ssk2831</td>
</tr>
<tr>
<td>93</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>9</td>
<td>2</td>
<td>10%</td>
<td>280</td>
<td>sll0707</td>
</tr>
<tr>
<td>94</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>95</td>
<td>Global nitrogen regulator</td>
<td>25046.3</td>
<td>6.53</td>
<td>7.2</td>
<td>6</td>
<td>38%</td>
<td>6.72E+03</td>
<td>sll1423</td>
</tr>
<tr>
<td>96</td>
<td>SOS function regulatory protein</td>
<td>22744.3</td>
<td>5.84</td>
<td>6.33</td>
<td>6</td>
<td>39%</td>
<td>1.28E+04</td>
<td>sll1626</td>
</tr>
<tr>
<td>97</td>
<td>PI protein</td>
<td>12397.4</td>
<td>6.33</td>
<td>6.96</td>
<td>6</td>
<td>58%</td>
<td>3.97E+03</td>
<td>sll0707</td>
</tr>
<tr>
<td>98</td>
<td>Photosystem I subunit IV linker protein</td>
<td>8145.2</td>
<td>8.03</td>
<td>8.82</td>
<td>4</td>
<td>78%</td>
<td>2.09E+02</td>
<td>ssk2831</td>
</tr>
<tr>
<td>99</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>100</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>101</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>102</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>103</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>104</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>105</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>106</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>107</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>108</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>109</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>110</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>111</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>112</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>113</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>114</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>115</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
<tr>
<td>116</td>
<td>Photosystem I subunit II</td>
<td>15643.9</td>
<td>8.95</td>
<td>8.76</td>
<td>2</td>
<td>36%</td>
<td>75.5</td>
<td>sll0707</td>
</tr>
</tbody>
</table>
Table 2. Isoforms of peripheral proteins

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>No. on 2D gel</th>
<th>Deduced pl</th>
<th>apparent pl</th>
<th>Cyanobase No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP synthase b subunit</td>
<td>4, 5, 6, 7, 8</td>
<td>4.89</td>
<td>4.89, 4.98, 5.01, 5.16, 5.25</td>
<td>slr1329</td>
</tr>
<tr>
<td>Protein translation elongation factor Tu</td>
<td>10, 11, 12, 13, 14, 15</td>
<td>5.16</td>
<td>5.25, 5.34, 5.43, 5.52, 5.61, 5.7</td>
<td>slr1099</td>
</tr>
<tr>
<td>ATP synthase a subunit</td>
<td>16, 17, 18, 19, 20</td>
<td>4.96</td>
<td>5.16, 5.20, 5.25, 5.34, 5.46</td>
<td>slr1326</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>26, 27</td>
<td>4.95</td>
<td>4.98, 5.16</td>
<td>slr0151</td>
</tr>
<tr>
<td>Plastoquinol-plastocyanin reductase</td>
<td>37, 38</td>
<td>4.87</td>
<td>4.5, 4.8</td>
<td>slr1316</td>
</tr>
<tr>
<td>12 KD extrinsic protein of PSII</td>
<td>40, 41</td>
<td>4.44</td>
<td>3.63, 4.26</td>
<td>slr1194</td>
</tr>
<tr>
<td>ATPase subunit epsilon</td>
<td>44, 45</td>
<td>5.2</td>
<td>5.61, 4.89</td>
<td>slr1350</td>
</tr>
<tr>
<td>PII protein</td>
<td>55, 56, 96</td>
<td>6.33</td>
<td>5.7, 5.97, 6.96</td>
<td>ssr3532</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>71, 72, 73</td>
<td>7.82</td>
<td>6.78, 7.14, 7.</td>
<td>slr1306</td>
</tr>
<tr>
<td>Light repressed protein</td>
<td>75, 76, 77, 84</td>
<td>6.07</td>
<td>6.78, 6.96, 6.82, 6.33</td>
<td>slr0947</td>
</tr>
<tr>
<td>Photosystem I subunit II</td>
<td>85, 86, 88, 89</td>
<td>8.95</td>
<td>9.6, 9.5, 9, 8.76</td>
<td>slr0737</td>
</tr>
</tbody>
</table>
Table 3. Many hypothetical gene-products associate with thylakoid membranes

<table>
<thead>
<tr>
<th>Protein</th>
<th>Transmembrane segments</th>
<th>Score for hydrophobic</th>
<th>Certainty</th>
<th>Size (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Helix</td>
<td>Begin</td>
<td>End</td>
<td></td>
</tr>
<tr>
<td>slr1306</td>
<td>1</td>
<td>6</td>
<td>26</td>
<td>1.094</td>
</tr>
<tr>
<td>slr1270</td>
<td>1</td>
<td>12</td>
<td>32</td>
<td>1.957</td>
</tr>
<tr>
<td>slr1835</td>
<td>1</td>
<td>16</td>
<td>36</td>
<td>1.985</td>
</tr>
<tr>
<td>slr0151</td>
<td>1</td>
<td>30</td>
<td>50</td>
<td>1.767</td>
</tr>
<tr>
<td>slr0431</td>
<td>1</td>
<td>20</td>
<td>40</td>
<td>1.579</td>
</tr>
<tr>
<td>slr1688</td>
<td>1</td>
<td>24</td>
<td>44</td>
<td>1.682</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>149</td>
<td>169</td>
<td>0.856</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>320</td>
<td>340</td>
<td>0.939</td>
</tr>
<tr>
<td>slr0503</td>
<td>3</td>
<td>359</td>
<td>379</td>
<td>1.218</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>382</td>
<td>402</td>
<td>1.591</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>405</td>
<td>425</td>
<td>0.76</td>
</tr>
<tr>
<td>Protein category</td>
<td>Protein No. on the 2D gel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Photosynthesis and respiration</td>
<td>4, 5, 6, 7, 8, 16, 17, 18, 19, 30, 37, 39, 40, 43, 54, 63, 69, 85, 86, 88, 89, 101, 103, 113</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cellular processes</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biosynthesis of cofactors, prosthetic groups, and carriers</td>
<td>9, 74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>10, 11, 12, 13, 14, 15, 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transport and binding protein</td>
<td>23, 33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulatory functions</td>
<td>78, 83, 90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fatty acid, phospholipid and sterol metabolism</td>
<td>44, 45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purines, pyrimidines, nucleosides, and nucleotides</td>
<td>60, 109</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical proteins</td>
<td>3, 21, 26, 27, 32, 34, 47, 58, 61, 62, 65, 71, 72, 73, 105, 106, 111, 112, 114, 115, 116</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other category</td>
<td>1, 75, 76, 77</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Effect of sample treatments on the recovery of integral and peripheral proteins of thylakoid membranes. All samples were loaded on an equal chlorophyll basis; the proteins in each lane were prepared from thylakoid membranes containing 10 μg chlorophyll. For (U/T 2h) and (U/T 0.5 h), the membranes were extracted with rehydration buffer (9M urea, 2% Triton-X-100) at room temperature for 2 and 0.5 h, respectively. The samples were centrifuged and proteins in the supernatant were used for electrophoresis. For (SDS 2h), (SDS 0.5h), the membranes were extracted with 1% SDS for 2 and 0.5 h, respectively. For TCA precipitation treatments, the membranes were incubated at room temperature with rehydration buffer or water for 2 h, then same volume of 10% TCA in acetone was added to samples, and proteins were precipitated for 2 h. The pellets were treated two times with acetone to remove residual pigments and lipids. The pellets were dissolved in rehydration buffer. In the control lane (The first lane from left), membranes incubated with denaturing dye (2% SDS and 0.1 M DTT) at 37 °C for 2 h were used.

Figure 2. Separation of the peripheral proteins of thylakoid membranes by 2D electrophoresis. The first dimension included 18 cm IPG strip (pH 3-10 nonlinear gradient), and the gradient gel (12-18%) was used for the second dimension. All labeled protein spots on the 2-D gel were analyzed with mass finger printing.
Figure 1
Figure 2
CHAPTER 4. THE PROTEOME OF CYANOBACTERIAL PHOTOSYNTHETIC MEMBRANES: IDENTIFICATION OF NOVEL PERIPHERAL AND INTEGRAL MEMBRANE PROTEINS

A paper to be submitted to Proteomics

Yingchun Wang*, Dan Harvey#, Suresh Kothari#, Parag R. Chitnis*2

Abstract

Prefractionation can help to resolve more membrane proteins in 2-D gels. In this study, we performed serial extraction to separate peripheral and integral thylakoid membrane proteins of *Synechocystis* sp. PCC6803 using 8 M urea. IPG strips with pH range 3-10, 4-7, 4-5 and 5-6 were used for the first dimension separation for peripheral proteins while pH 3-10 and pH 4-7 were used for integral proteins. In the peripheral fraction, 249 out of 310 analyzed protein spots were identified as expression products of 95 individual genes, of which 61 proteins were not identified in our previous study. In the integral fraction, 63 out of 132 analyzed spots were identified as 29 different proteins, of which 13 proteins were not identified either in peripheral fraction or in our previous study. Totally, 128 individual proteins were identified from both fractions including proteins identified in our previous study, of which 41 proteins have been identified in Cyano2Dbase including 13 proteins that were classified as thylakoid proteins. Among the new identified proteins, 38 are encoded by hypothetical genes, 13 of which contain known functional motif and domain that found in other proteins. Hydropathy analysis shows that 24 proteins from the integral fraction contain at least one transmembrane helix, including all of the 17 identified hypothetical proteins.

1. Introduction

Two-dimensional gel electrophoresis in conjunction with mass spectrometry has been widely used for global analysis of proteins. However, membrane proteins remain a major

---

1 This work is supported in part by grants from Iowa Corn Promotion Board and from the Vice-Provost for Research at ISU. Journal Paper No. J-XXXX of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3416 and supported in part by Hatch Act and State of Iowa funds.

2 * Department of Biochemistry, Biophysics and Molecular Biology, # Department of electrical and computer engineering, Iowa State University.
challenge for 2-DE analysis due to their insolubility and low abundance. Fractionation of protein samples is one of the best ways to solubilize more hydrophobic, lower abundant proteins. It has been demonstrated that sample fractionation can help to solubilize more integral proteins. In addition, sample fractionation can help to localize sub-cellular location of a protein and relate the function of a protein to specific sub-cellular compartments.

The thylakoid membranes of cyanobacteria contain many proteins involved in photosynthesis and respiration; the two essential physiological pathways for cell growth. In addition, thylakoid membrane proteins may play critical roles in membrane biogenesis, biosynthetic pathways, membrane transport and regulation of energy metabolism. Therefore, functional analysis of these proteins will be necessary for answering many genetic and physiological questions.

The unicellular cyanobacterium strain *Synechocystis* sp. PCC 6803 has been widely used as a model system for molecular genetic, biochemical and biophysical studies of photosynthesis and other chloroplast-related functions [1]. It is the first oxygenic photosynthetic organism of which the genome has been completely sequenced [2]. The 3.57 Mb-genome contains 3168 ORFs, about 50 % are assumed to encode putative proteins with unknown functions, and many of which have homologs founded in chloroplast of high plants. Proteome analysis of the organism has been conducted with 2-DE in different groups recently [3-6]. In one group, N-terminal microsequencing was adopted to identify proteins resolved in 2-D gels. Of the 244 spots analyzed, the N-terminal sequences of 234 spots have been successfully determined, and 227 of which were confirmed to be encoded by 143 independent corresponding genes in the genome. Out of these genes 34 produced multi-spots proteins with different pI, and 14 genes encode proteins associating with thylakoid membranes [4,5]. The other group resolved pure plasma membrane proteins in 2-D gels and identified 57 different proteins, of which 17 were predicted to have one or more transmembrane helices [3]. Our previous proteomic study of thylakoid membrane associated proteins has identified 78 protein spots derived from 51 independent genes, including the products of 17 hypothetical genes, and 8 genes produce multi-spots of polypeptides [6].

In the present study, we adopted a sample fractionation approach using 8 M urea to separate peripheral and integral thylakoid proteins. A new IEF solution including Sulfiobetain
(SB-10) as well as narrow pH range IPG strips were used in order to resolve more proteins. Ninety five and 29 different proteins were identified from each fraction respectively. Hydropathy analysis was also performed for proteins from both fractions.

2 Materials and Methods

2.1 Growth of *Synechocystis* sp. PCC 6803 and preparation of membrane

The wild-type strain of *Synechocystis* sp. PCC6803 was grown in BG11 medium with 5 mM glucose under 40 μmol.m⁻².s⁻¹ illuminations at 30 °C. For membrane preparation, cells at a late exponential phase were harvested and resuspended in a buffer (0.4 M sucrose, 50 mM MOPS, pH 7.0, 10 mM NaCl, 5mM EDTA, and 0.5 mM PMSF). Cells were broken using a bead beater and thylakoid membranes were isolated by differential centrifugation [7]. Chlorophyll concentration of the thylakoid preparation was measured in 80% acetone using a UV-160U spectrophotometer (Shimadzu) [8].

2.2 Isolation of peripheral and integral thylakoid membrane proteins

The purple phycobilosomes in cytoplasm of *Synechocystis* were used as visual indicators of cytoplasmic contamination of the membranes. Thylakoid membranes were washed with 20 mM MOPS, pH 7.0, 50 mM EDTA for 2 or more times, until the washing buffer was colorless. Purified membranes were extracted with 8 M urea to release the peripheral proteins, and then centrifuged at 75, 600 x g to pellet the insoluble fraction that was enriched with integral membrane proteins. The pellet was extracted with 8 M urea again, and the supernatants of the two extractions were combined and labeled as peripheral fraction. Similarly, the pellets from two consecutive extraction steps were pooled for analysis as the integral membrane fraction. The peripheral fraction was diluted four-times with deionized water and centrifuged at 75, 600 x g to pellet the carry over insoluble fraction which was combined into the integral membrane fraction. The peripheral fraction was then precipitated with 10% TCA on ice for 30 min prior to extraction with 100% ice-cold acetone to remove lipids and pigments. The integral fraction was also washed with ice-cold acetone multiple times until the washing acetone was colorless. Both fractions were dried under vacuum, dissolved with multiple surfactant solution (5 M urea, 2 M thiourea, 2 mM TBP, 2% CHAPS,
2% Sulfobetaine 3-10, 0.5% carrier ampholytes, 40 mM Tris, 0.001% orange G dye), and sonicated for 15 minutes in a water bath. At this step, almost all peripheral proteins can be dissolved, whereas the integral proteins can only partially be solubilized. The insoluble parts in both fractions were removed by centrifugation at 75,600 x g. Protein concentration of both fractions was measured with Bio-Rad DC Protein assay kit (Bio-Rad, Richmond, CA, USA).

2.3 2-DE of peripheral proteins of thylakoid membranes

The immobilized pH gradient strips with different pH ranges (18 cm, pH 3-10, non-linear, pH 4-7, pH 4-5, pH 5-6, Pharmacia Biotech, Uppsala, Sweden) were rehydrated by 320 µl sample solutions containing about 500 µg proteins from corresponding fractions (For the gel stained with silver nitrate, about 80 µg proteins were loaded on the IPG strips with pH 3-10). Active rehydration was performed by applying low voltage (20 V) for 10 hours after 2-hour rehydration without voltage at 20 °C. The first dimensional IEF with pH 3-10 and pH 4-7 were performed with an IPGphor instrument (Pharmacia) with the following voltage settings: 100 V for 0.5 h, 300 V for 0.5 hr, 1000 V for 0.5 h, 2500 V for 0.5 h, 5000 V for 0.5 h, and then 8000 V until a total of 80 000Vh was reached. For narrow pH range IPG (pH 4-5 and pH 5-6), the settings were the same except that a total of 12, 000 Vh should be reached. Upon electrophoresis, the proteins on the strips were denatured and cysteinyl residues were reduced by equilibrating the IPG strips with a buffer containing 6 M urea, 2% SDS, 0.375 M Tris/HCl, pH 8.8, 20% glycerol, 5 mM TBP, 2.5 % acrylamide monomer for 20 minutes. The second-dimensional electrophoresis was performed using 12-18% gradient SDS-PAGE gels. Upon electrophoresis, the protein spots on the SDS-PAGE gel were stained with colloidal Coomassie blue and the gels were scanned with GS-710 Calibrated Imaging densitometer (Bio-RAD) to obtain images for analysis with Melanie II software [9].

2.4 Mass spectrometry

Protein spots that were visualized with colloidal Coomassie blue were excised manually and incubated at 37°C with 2.5 mM Tris HCl (pH 8.5) in 50 % acetonitrile to remove the dye bound to the protein. The gel pieces were dried under vacuum followed by incubation with 10 µl of 10µg/ml trypsin in 2.5 mM Tris-HCl, pH 8.5 at 37 °C for 18 h. The resulting tryptic fragments were eluted by diffusion into 50% acetonitrile and 0.5%
trifluoroacetic acid (TFA). Diffusion of peptide fragments was facilitated with ultrasonication in a water bath at 4°C. One micro liter of the tryptic peptides of each sample was mixed with 1 μl α-cyano-4-hydroxy-cinnamic acid matrix prior to be transferred to a 100 well plate for MALDI-TOF. A Voyager-DE™ PRO Biospectrometry Workstation was used to acquire mass spectra in a reflection-delayed extraction mode over a mass range of 600-4000 Da. The final mass spectra were the accumulation of the spectra obtained from 3-6 positions with 64 shots (total 192-384 shots). If high resolved mass spectra could not be obtained for a spot, the sample would be concentrated by drying again with vacuum and resuspended in 2.5 μl of elution buffer and 2.5 μl of 2.5 mM Tris HCl (pH 8.5). Increased concentration of peptides could produce mass spectra for almost all protein spots that could be visualized with colloidal Coomassie blue. Some proteins may have less tryptic sites and cannot produce enough tryptic peptides within the mass range 600-4000 Da for identification purpose. Therefore, spectra within mass range 3000-6000 Da were also acquired to ensure good resolution for the spectra between 3000-5000 Da. Subsequently, the peptide mass fingerprints (PMF) generated from the two mass ranges were combined as one PMF to search the database. The peptide ions generated by autolysis of trypsin (with m/z 832.33+, 842.51+, 1045.56+, 2211.10+) were used as internal standard peaks for calibration. The mass spectra were analyzed and the PMF for each sample was generated with the Data Explorer™ software.

2.5 Data analysis

Peptide masses were analyzed with the MS-Fit program (http://prospector.ucsf.edu) using the following parameters: mass tolerance of 25 ppm, a minimum of four peptide matches and with 1 missed cleavages. The database searching was performed using an automated MS-FIT program. The motif and domain of hypothetical proteins was predicted by InterPro program (http://www.expasy.ch). Hydropathy analysis for deduced protein sequence was performed by predicting transmembrane helices (TM) using TopPred2 program (http://www.expasy.ch) and by calculating the GRAVY value [10]. The prediction of transmembrane helices for the whole proteome was performed using an automated TMpred program [11]. The homology analysis was performed using BLAST program.
3 Results

3.1 Sample preparation from membranes

Membrane proteins are generally refractory to aqueous solutions and hard to be resolved by 2-DE. In our previous work, we identified 58 peripheral proteins from thylakoid membranes of *Synechocystis* sp. PCC6803 [6]. In order to resolve more membrane proteins in 2-D gels, we performed serial extraction to separate peripheral from integral proteins. Figure 1 shows a colloidal Coomassie blue stained SDS-PAGE gel with protein samples from different fractions. As a control, thylakoid membrane was denatured with SDS and loaded in the first lane. As we described before, many thylakoid specific proteins such as PsaAB, CP47, PsaD and PsaC can be identified from the apparent migration [6]. The second lane is the peripheral proteins extracted with 8 M urea. In this fraction, most of the peripheral proteins were recovered and enriched and most of the integral proteins are missed. For examples, the bands of known peripheral proteins PsaD and PsaC are present in this lane with stronger intensity while the dominant band for the 68 kD integral protein PsaAB is missing. In the lane for the integral fraction, less protein bands are present, including some known carry over peripheral proteins such as PsaD. The pattern of the bands in the integral fraction is different from that of peripheral fraction, indicating that the fractionation has been successful. However, the integral protein PsaAB is still not observed in the integral fraction. We tried to use 2% SDS to solubilize the insoluble part from the integral fraction that was extracted with multiple surfactant solution, and still cannot observe PsaAB on SDS-PAGE gels (Data not shown). Therefore, it is obvious that the loss of PsaAB is due to irreversible precipitation caused by TCA. It seems that it is inevitable that some integral proteins will lost for preparations using TCA. Some protein bands in integral fraction are not visible on the total membrane fraction, such as the dominant band formed at about 50 kDa. This could be explained by the protein enrichment caused by fractionation. These proteins are present at low abundance in thylakoid membrane and shielded by the spots of other high abundant proteins in SDS-PAGE gel, but are concentrated during the sample fractionation. Both the peripheral and integral fractions contain many high molecular weight proteins that are lost in our previous extraction [6]. It is obvious that the serial extraction with 8 M urea followed by TCA/acetone precipitation and acetone washing is more advantageous than the direct
membrane precipitation with TCA in improving protein number and quantity. Therefore, we used this strategy for our routine protein preparation for 2-DE analysis of thylakoid membrane proteins.

3.2 2-D separation of peripheral proteins

Proteins from both fractions were separated according to their pH using IEF and further resolved according to their mass with SDS-PAGE. IPG strips with pH range 3-10 (nonlinear) and 4-7 were used for separation. Because a large number of peripheral proteins solubilized and most of them are within pH 4-7 [6], we also tried to separate peripheral proteins with narrow pH range such as 4-5 and 5-6. The IPG strips with pH range 6-11 were also tried to separate basic proteins, but the resolution was not good due to horizontal streaking caused by electroendosmosis [12].

Figure 2 shows the 2-D separation of peripheral proteins with different pH range for the first dimension. Fig. 2a is a silver stained 2-D gel with pH range 3-10 (nonlinear), about 400 protein spots were detected. As we can see from the gel map, most of the protein spots are within the pH range 4-7. From a heavy loaded parallel-run 2-D gel stained with colloidal Coomassie blue, we excised 120 protein spots for further analysis. Fig. 2b is a 2-DE gel with pH range 4-7 stained with colloidal Coomassie blue; about 240 protein spots are detected. Since colloidal Coomassie blue is not as sensitive as silver staining, less number of protein spots are expected to be detected in this gel than the number of spots detected in the region of pH 4-7 in the gel of pH 3-10 stained with silver nitrate. We excised 110 spots from this gel for identification. Fig. 2c and 2d are 2-DE gels with pH range 4-5 and 5-6 respectively, both are stained with colloidal Coomassie blue, about 40 and 100 protein spots are detected on each gel respectively. We excised 32 spots from the gel with pH 4-5 and 48 spots from the gel 5-6 for further analysis. All the excised protein spots were digested with trypsin and the resulting peptides were subjected to PMF analysis using MALDI-TOF mass spectrometry.

Proteins were identified in the same way as we described before [6]. The identity of proteins with high scores in the MS-FIT analysis was further validated with three other criteria: mass, pH and corresponding to the spots in Cyano2Dbase. The deduced mass of the putatively identified proteins should match the protein mass estimated from the migration on the 2-D gel. In the mass comparisons, we considered possible post-translational processing
which could increase or decrease the mass. The second criterion was the deduced pI of proteins, which should be close to the pI estimated from the 2-D gel. Again, the possibility of post-translational modification that may change pI was considered. The last criterion is the correspondence to the identified proteins in the Cyano2Dbase. If a protein in our gel was identified as the same protein in the Cyano2Dbase and was located at the same place in a gel, then the identification is considered to be correct. Some identity of known proteins with low Mascot score was further validated by N-terminal sequencing and electroblotting, such as PsaC and PsaD, and the information of these proteins were also included in the tables for identified proteins.

In the 2-D gel with pH range 3-10 (P 3-10), 89 spots have been identified as protein products encoded by 57 independent genes. Some previously identified dominant protein spots such as ATP synthase β subunit and the proteins synthesis elongation factor Tu were not excised for PMF analysis (Fig. 2a). Seventeen proteins were observed to have isoforms that differentiate with each other in pI (Table 1), indicating that post-translational modification leading to pI change are very common for thylakoid proteins. Many known peripheral proteins such as PsaD (p75 in table 1) and PsaE (p94 in table 1) exist as multiple or dominant spots in the 2-D gel. Some soluble proteins that function through association with membrane are also present in the peripheral fraction. For example, the phycocyanin a and b subunits (p69, p67 in table 1) are two proteins of phycobilosomes that are known to associate with the PSII complexes in the thylakoid membranes [13]. Some proteins appeared not only at expected molecular mass value, but at multiples thereof. For example, the 16 kDa protein PsaD appeared as 7 individual spots, 6 of them located at the places where the mass value is 16 kDa, the exception is the spot 64 that locate at the place where the mass value is about 32 kDa (Fig. 2a). It is obvious that it is a homodimer formed during electrophoresis due to the regeneration of disulfide bridges caused by insufficient alkylation of proteins [14]. We also found heterodimers or oligomers in our past proteomic study (Data not shown). Therefore, alkylation step should be included before 2-DE for the later 2-D analysis [14].

We identified 99 spots encoded by 59 independent genes from the gel with pH 4-7 (P 4-7) (Table 1 and Fig. 2b), of which, 33 independent gene encoding products were not identified in the gel P 3-10. Almost all of the new identified proteins are low abundant
proteins. For example, both of the protein NADH dehydrogenase subunit 7 (p10 in Table 1) and Phosphate transport ATP-binding protein PstB (p40 in Table 1) have several isomers existing in low abundance. The low abundance of these proteins makes them either under the detecting range of the employed staining methods or shielded by other high abundant proteins. Use of narrower pH range gel can allow more sample loading and higher resolution. Therefore, more low abundant proteins are expected to be detected and identified in the same pH range for narrower pH range 2-D gels. Based on this idea, we also tried to separate peripheral proteins using 2-D gels with pH range 4-5 (P 4.5) and 5-6 (P 5.5). In the gel P 4.5, 21 spots were identified as protein products encoded by 13 independent genes, and 4 of the independent gene encoded products were not identified in the gel with pH range 3-10 and 4-7. In the gel P 5.5, 40 spots were identified as protein products encoded by 24 independent genes, and also 4 independent gene encoded products were not identified in the gel with pH range 3-10 and 4-7 (Table 1).

Totally, 249 spots out of the analyzed 310 spots from the four 2-D gels have been identified (Table 1). The rate of identification is as high as 80%, which is much higher than the 67% in our previous study [6]. The increase of identification rate is mostly due to the sample concentrating step we used for low abundant proteins. Many low abundant proteins that cannot produce mass spectra with high resolution can produce spectra that are good enough for PMF analysis after concentrating. The remaining spots could not be identified because of the poor quality of samples, lower amount of proteins, less number of unmodified peptides in these proteins or errors in calibration. The 249 proteins identified spots are encoded by 95 independent genes (Table 1). In our previous study, protein products of 51 independent genes have been identified, 20 of them were not identified here. Therefore, a total of 115 individual gene products have been identified in the peripheral fraction.

3.3 2-D separation of integral membrane proteins

Integral proteins were separated in 2-D gels with pH range 3-10 (I 3-10) and pH 4-7 (I 4-7) in the same way as peripheral proteins (Fig. 3). About 120 spots were observed in the gel I 3-10 and 80 spots were observed in the gel I 4-7. Sixty five and 67 spots from each gel were subjected for PMF analysis, respectively. Thirty spots were identified as protein products...
encoded by 17 individual genes in the gel I 3.10 while 33 spots were identified in the gel I 4.7 (Table 1). The rate of identification is only 48%, which is much lower than that of peripheral proteins. We assume that the low identification rate may be caused by more post-translational modification sites on integral proteins. Totally, 63 protein spots were identified as expression products of 29 individual genes. Table 1 shows that 11 of these genes are hypothetical; and 11 gene products appear as multiple spots. Among the 29 individual proteins, 17 proteins are specific to integral fraction and the other 12 proteins also appear in peripheral fraction. The co-existence in both fractions of thylakoid proteins can be caused by two factors: One is that 8 M urea is not powerful enough to extract all peripheral proteins from membrane, which leads to the carry-over contaminants of peripheral proteins in integral fraction. For example, the protein PS I subunit I (PsaD) are known peripheral proteins, but it can be found in both fractions (p75). The other factor exists during the fractionation step for sample preparation. The resuspension of thylakoid membrane in 8 M urea makes it very difficult to pellet the insoluble fraction (integral fraction). Even though we diluted the solution 4 times, there were still some green membranes exist in the extracted peripheral fraction (supernatant). Therefore, it is inevitable that some integral proteins present on peripheral fractions. For example, the 85 kDa general secretion pathway protein D are predicted to contain 3 transmembrane domains (Table 2), it should be classified as integral proteins but appeared in both fractions (p03). The contamination could be decreased by increase the time and speed for centrifugation during sample preparation. Topology analysis shows that 24 individual proteins in integral fraction contain one or more TMs, including the two proteins that contain 5 TMs (i03, i04). The presence of the other 5 proteins that do not contain TM in integral fraction may due to the reason we just discussed.

3.4 Summary of identification of thylakoid proteins

Table 4 shows the summary of our protein identification. In our present research, a total of 442 protein spots were analyzed, of which 312 protein spots were identified as protein products of 112 individual genes (Table 1). The overall identification rate is 70% percent. As shown in our previous study, multi-spots proteins are very common and a total of 45 proteins with multi-isoforms were observed. Of the 51 previously identified proteins, 16 were not identified here. Thus a total of 128 individual proteins were identified (Table 4),
including 41 proteins that are also identified in Cyano2Dbase [4,5]. Of the 14 identified thylakoid proteins in Cyano2Dbase, 13 were identified in our research. It is obvious that our study is more thylakoid specific and contains more thylakoid membrane proteins than Cyano2Dbase does.

### 3.5 Products of hypothetical genes

The genome of *Synechocystis* contains 50% (1562) hypothetical genes that encode proteins with unknown functions. In our previous study, we have identified expressed products of 17 individual hypothetical genes [6]. The current study has identified 25 and 11 hypothetical proteins from peripheral and integral fractions, respectively. Of the 11 hypothetical proteins from integral fraction, 6 proteins are specific to integral fractions and 5 proteins appeared in both fractions (Table 1). Including the 7 hypothetical proteins identified before that are not identified here, totally 38 hypothetical proteins were identified (Table 4). Most hypothetical proteins identified are low abundant, which may be one of the reasons that their functional analysis lagged behind other proteins. Several hypothetical proteins are abundant proteins forming the major spots in 2-D gels with several isoforms. For example, the hypothetical protein encoded by the ORF slr1506 formed highly concentrated protein spots at 68 kDa with 4 isoforms in the 2-D gel I3-10 (i01), the unknown function of these abundant hypothetical proteins may due to their insolubility in normal aqueous solutions, which hampers their functional analysis.

The topology analysis for deduced amino acid sequences of hypothetical proteins has been performed using TopPred2 (Table 2). Table 2 shows that 10 out of the 25 identified hypothetical proteins from peripheral proteins contain one or two transmembrane helices including 5 hypothetical proteins that also appear in integral fraction. Almost all of the putative transmembrane helices are located near the N-terminus and thus assumed to be part of the lumen-targeting presequences [15], the only exception is the protein encoded by ORF slr1275, which have another one less-hydrophobic region well inside the sequence, we assume that the less hydrophobic region may indeed be used for interaction with the hydrophobic lipid bilayer rather than for crossing the membrane. All the other hypothetical proteins that do not contain predicted TM may interact with membrane through the way other than hydrophobic interaction, such as ionic interaction or hydrogen bonding. Table 2 shows
that all hypothetical proteins identified from integral fraction contain one or more TMs. The proteins encoded by ORF slr1506 and sll1307 contain at least one TM that locates well inside the protein sequence, these proteins can be considered as integral proteins with no doubt. All other proteins that contain one hydrophobic region are either real integral proteins or peripheral proteins that have strong interaction with lipid bilayer.

To investigate the function of all hypothetical proteins is a heavy task for all researchers. However, the intracellular location of the hypothetical proteins could help us to understand their functions. For example, all the hypothetical proteins here are related with thylakoid membrane, their function are therefore related with either photosynthesis, respiration or other functions carried out by thylakoids. Table 3 shows that 13 hypothetical proteins contain one or more domains and motifs predicted by the program InterProScan (http://www.expasy.ch ). Functions of hypothetical proteins can be predicted from the functional domain and motif they contain. For examples, the proteins encoded by ORF slr1506 have two domains; the esterase/lipase/thioesterase domain indicates that the protein is a proteolytic enzyme exploit serine in their catalytic activity that is ubiquitous in viruses, bacteria and eukaryotes; the domain ATP/GTP-binding site motif A indicates that the proteolytic activity might need energy from ATP or GTP. The protein encoded by ORF slr0038 belongs to mitochondrial energy transfer proteins (carrier proteins) family that are found in the inner mitochondrial membrane, such proteins include: ADP,ATP carrier protein (ADP/ATP translocase); 2-oxoglutarate/malate carrier protein; phosphate carrier protein; tricarboxylate transport protein (or citrate transport protein); Graves disease carrier protein; yeast mitochondrial proteins MRS3 and MRS4; yeast mitochondrial FAD carrier protein; and many others [16-20]. In Synechocystis, the thylakoid membrane contains machinery for energy metabolism such as photosynthesis and respiration electron transport chains [21]. Therefore, it is highly possible that the hypothetical protein encoded by slr0038 is a substrate carrier protein functionally related with thylakoids.

Functions of hypothetical proteins can also be deduced from homology analysis, that is, a hypothetical protein may carry the same function as its homologous proteins with known function. However, homologs of all hypothetical proteins identified by us are also have unknown functions, thus it is hard to obtain the functional information of these proteins at
current time. Fortunately, with more and more proteins studied world-wide, the databases are updated very frequently. As an example, 50 hypothetical proteins was identified from our local databases initially, and 38 hypothetical proteins were corroborated with the updated NCBI databases and Cyanobase database, the function of the rest 12 proteins are now unraveled. Therefore, it is possible that functions of the hypothetical proteins or their homologs will be known in the near future.

4 Discussion

Membrane proteins associate with a lipid bilayer in several ways. Integral or intrinsic proteins cross the lipid bilayer for one or more times with transmembrane helices consisting of dominant hydrophobic amino acid residues. Peripheral proteins do not cross the lipid bilayer, but associate with membrane through multiple types of interactions as reviewed by Santoni et al. [22]. Basically, the association can be classified as covalent or non-covalent. Covalent association can be mediated through a fatty acid, a polyisoprenyl chain or glycosylphosphatidyl inositol (GPI) anchor [23,24]. Non-covalent association is achieved through hydrophobic interaction, ionic interaction or hydrogen bonding. Covalently associated peripheral proteins can be wiped out of membrane by specific enzymes breaking the bond linking membrane and the proteins. As an example, the GPI-specific phospholipase C has been used successfully to release GPI-proteins from membrane [23,24]. Non-covalent associated peripheral proteins can be released from membrane with multiple approaches. For examples, chaotropes such as urea can be used to interrupt hydrogen bonding [25,26], high pH solution such as sodium carbonate can be used to disrupt ionic interaction [27,28], and organic solvents can be used to disrupt hydrophobic interactions [29].

In order to fractionate peripheral and integral thylakoid proteins of *Synechocystis*, we tried to use the chaotrope urea to extract peripheral proteins from membrane. The vast majority of known proteins in peripheral fraction are soluble, and most hypothetical proteins contain no or only one TM. In contrast, all hypothetical proteins in integral fraction contain one or more TMs, only five known proteins do not contain TM. It is obvious that the fractionation has been successful. In order to further validate our hydropathy analysis, we also calculated GRAVY value for proteins from both fractions [10] (Fig. 4). Fig. 4a and 4b shows the distribution of GRAVY for peripheral and integral proteins, respectively. All the
identified peripheral proteins are within GRAVY -1.0 – 0.4, most of them are within -0.4 - 0.2. Similarly, all the integral proteins are within GRAVY -0.8 – 0.6, and most of them are within -0.4 – 0.2. Interestingly, the vast majority of both peripheral and integral proteins have negative GRAVY value, the result is agree with the results reviewed by Santoni at al [22], where all integral membrane proteins from bacteria identified in 2-D gels are negative in GRAVY. The GRAVY distribution pattern for both fractions is very similar to the distribution for all the proteins in the proteome of Synechocystis (Fig. 4c). Fig. 4c shows that the vast majority of proteins in the proteome of Synechocystis are within the GRAVY -0.6-0.2, which is the range that the GRAVY score of most of the thylakoid membrane proteins fall in. The highly negative proteins (GRAVY < -1) are supposed to be cytoplasmic proteins that were removed from thylakoids during membrane preparation, whereas the highly positive proteins (GRAVY >0.2) are expected to be integral proteins that are hard to be solubilized by current solution used for 2-DE.

The number of TMs and the value of GRAVY are two widely adopted criteria for evaluating the hydropathy of membrane proteins. Both criteria are not sufficient if used along. For example, 19 identified proteins from peripheral fraction have positive GRAVY value, whereas 26 identified proteins from integral fractions have negative GRAVY value (Fig. 4a, 4b), including the 5 TMs containing protein encoded by ORF sll0772 (-0.15). The contradiction could be explained as such that some big integral proteins not only contain multiple TMs, but also contain a large fraction of extra-membrane located loops; high contents of hydrophilic side chains of amino acid residues within the loops decrease the overall GRAVY value of the proteins. In contrast, small peripheral proteins some time contain one or more hydrophobic region that are used to interact with integral proteins; high contents of hydrophobic side chains of the amino acids within the hydrophobic region can increase the overall GRAVY value of the protein. Furthermore, some peripheral proteins may have one putative TM at N-terminus that may be used for targeting the protein to thylakoid lumen, just as we described in previous paragraphs, we cannot classify this protein as integral protein based only on the number of TMs. Therefore, we must be cautious when evaluating hydropathy for membrane proteins. In order to investigate the relationship between GRAVY value and the number of TM, we plot the number of TM against the GRAVY value for all
proteins from the proteome of Synechocystis (Fig. 5). As we expected, most of the proteins that do not contain TM have negative GRAVY value, and less proteins have positive GRAVY. The percentage of proteins with negative GRAVY value increased for the proteins with 1 TM. An apparent trend can be observed that the percentage increased with the increase of the number of TM. For proteins with 5 or more TMs, only one of them has negative GRAVY value. Again, the highest GRAVY does not come from the protein containing the most number of TMs, but from a protein contains 3 TMs (Fig. 5).

Analysis of membrane proteins remains a major challenge for proteomics techniques based on 2-DE because of poor solubility of membrane proteins in aqueous solutions. In our study, two membrane proteins containing 5 putative TMs encoded by ORF sll0072 and slr0891 respectively have been resolved on 2-D gels (Table 1). In the other report for plasma membrane proteome, a protein containing 3 putative TMs encoded by ORF slr0677 was identified from 2-D gel [3]. No proteins containing more than 5 TMs has been identified in Cyano2Dbase [5]. Even though our current approaches are more efficient in resolving hydrophobic proteins, proteins with more TMs such as PsaAB, which are high abundant and contains 11 TMs, is still lost during 2-D procedure.

It has been demonstrated that the insolubility of membrane proteins comes from the poor efficiency of most commercial detergents to solubilize denatured membrane proteins under the conditions prevailing in 2-DE [30]. To date, the most powerful detergent used is SDS; it can solubilize almost any proteins with the concentration of 2%, including the protein PsaAB (Fig. 1). However, SDS is highly negatively charged and cannot be used for isoelectric focusing, which requires low salt and electrically neutral detergent conditions. Although some protocol suggested that SDS could be used to extract proteins from membrane followed by extensive dilution with non-ionic or zwitterionic detergents [25,31,32], it is not applicable when high load samples for IEF are required. Some new zwitterionic detergents have been synthesized to enhance the solubilizing power of standard urea/chaps solution [33,34], and the human red blood cell plasma membrane proteins Band III, which has 12 TMs, can be solubilized with new synthesized zwitterionic detergent [30]. But unfortunately, these detergents are still not commercial available, and each of the solutions containing a different new synthesized detergent can just solubilized a limited kinds
of membrane proteins [27,30,33,35]. Therefore, a new powerful detergent that can be used for solubilize wider range of membrane proteins should be developed in order to resolve more hydrophobic membrane proteins. The proteins with extreme pI are also underrepresented in the current proteome databases. It has been estimated that about 20% ORFs of an organism encoding proteins with pI higher than 10.0 [36], and most of the integral membrane proteins have an alkaline pI [37]. In our research, the protein with the highest pI identified so far is 50s ribosomal protein L5, which has a deduced pI 9.69 (Table 1), no protein with pI higher than 10 identified in cyano2Dbase or in the other report for plasma membrane proteome [3]. Although extended pH gradients such as pH 6-11 are also available commercially, the use of this kind of IPG strips brings some new problems. Horizontal streaking is very common in 2-D gels with pH range 6-11 because of endosmosis effects [12]. Fortunately, more efforts have been spend on 2-D separation of alkaline proteins and better 2-D patterns have been obtained recently [38-40]. In addition, proteins tend to precipitate in IPG at the place where the local pH is equal to their pI [22]. In our study, we observed that the precipitation produced many wide bands on the strips, even the use of extended time for the equilibration of IPG strips could not transfer them to the second dimensional SDS-PAGE gel (data not shown).

The hydrophobic interaction between proteins and lipids is also one of the major factors that cause the insolubility of some integral membrane proteins, and large amount of lipids in samples could block gel pores of IPGs and cause smears and horizontal streaking on the subsequent second dimension gels. Although it has been said that some detergent works well for delipidated samples and some for samples without delipidation [22], our results shows that samples delipidated with acetone always provides the best results (data not shown), and this is consistent with the results reported by Simoes-Barbosa et al. [41].

High abundant proteins resolved in 2-D gels can generate highly resolved mass spectra that are good for PMF analysis, as the proteins ATP synthase b subunit and the protein synthesis elongation factor Tu always does, whereas low abundant proteins often have trouble in generating spectra with signal strong enough for PMF analysis. To solve this problem, we adopted a simple but effective method to concentrate samples, just as described in the material and methods part. According to our experience, if the protein spots stained
with CBB can be visualized with naked eye, then the concentrated sample of tryptic peptides can produce mass spectra with signal strong enough for PMF analysis. The concentrating method is not useful for protein spots that can only be detected with silver staining, because these proteins generally have much lower abundance. For such proteins, the most effective and simplest way to increase their concentration is to combine the same protein spots from several 2-D gels run simultaneously.

Fractionation of protein samples can also enrich low abundant proteins [42], this can be corroborated by the fact that the total number of proteins we identified from peripheral and integral fraction is much bigger than number we identified before [6]. We must be conscious that the some proteins we identified here or before might not related with thylakoid membrane, but related with plasma membrane, because the two membrane fractions were not separated during membrane preparation. Even though thylakoid membrane is dominant in Synechocystis cells, carry-over contaminants of plasma membrane proteins are inevitable. In order to obtain purified membrane, polymer two-phase portioning can be adopted to fractionate thylakoid from plasma membrane [3,43]. In a recent report from Huang et al [3], purified plasma membrane proteins have been resolved and identified from 2-D gels, 57 different membrane proteins were identified, of which 17 were predicted to have one or more TMs. Interestingly, 29 proteins identified in that report also be identified in our study, including many known thylakoid proteins such as PsaC, PsaD and ATP synthase b subunit. Some of these proteins such as proteins encoded by ORF slr1908, slr0447 are real plasma proteins, we identified them here because of contamination, just as we discussed. But the presence of many known thylakoid proteins in plasma proteins was explained as such that plasma membranes are the initial site for photosystem biogenesis [3,44]. Thylakoid membrane proteins can be further fractionated using sucrose gradient centrifugation. In one report from our group, the hydrophobic protein PsaF, which contains 2 TMs and was not identified here, was resolved on the 2-D gels for purified photosystem I subunits [45].

A proteome contains more information than a genome does. For example, from the isoforms of PsaD (p75 in table 1), we know that this protein is post-translationally modified such that its pI is altered. This kind of information could not be obtained from genome sequence. Two-dimensional electrophoresis in conjunction with mass spectrometry has been
proven to be one of the most powerful tools for proteomic analysis because of its high
capacity for resolving thousands of proteins on a single 2-D gel [46]. Although there are many
problems that may hinder the wider-use of 2-DE for proteome analysis, the development in
this area have been immense in the past a few years. In order to use 2-DE as the routine tool
for large scale proteome analysis, more effort must be spent in the following area. 1) More
powerful nonionic or zwitterionic detergent must be invented; 2) Fractionation should be
widely adopted for sample preparation; 3) Narrow pH range IPG strips must be adopted as
routine tool for protein separation; 4) More sensitive staining technology must be employed;
5) More sensitive and higher resolution mass spectrometry should be used.

Acknowledgements
We thank Dr. Wu Xu for assistance with 2-DE and Dr. Jochen Kruip for providing
protocol of sample fractionation. This work is supported in part by grants from Iowa Corn
Promotion Board and from the Vice-Provost for Research at ISU. Journal Paper No. J-
XXXXX of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa,
Project No. 3416 and supported in part by Hatch Act and State of Iowa funds

References
956-66.
837-44.
[40] Chan, C.W., Warlow, R.S., Chapuis, P.H., Newland, R.C. et al. 1999 Electrophoresis 20, 3467-71. [pii].
Table 1. Identification of peripheral and integral thylakoid proteins

<table>
<thead>
<tr>
<th>No.</th>
<th>Protein ID</th>
<th>Label of identified proteins on 2-D gels with different pH range</th>
<th>Deduced mass (Da)</th>
<th>Deduced pi</th>
<th>Cyanobase No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hypothetical protein</td>
<td>6, 7, 8, 9</td>
<td>68847.7</td>
<td>5.16</td>
<td>slr1506</td>
</tr>
<tr>
<td>2</td>
<td>Probable porin; major outer membrane protein</td>
<td>4, 7, 8, 9</td>
<td>64511.2</td>
<td>5.07</td>
<td>slr1908</td>
</tr>
<tr>
<td>3</td>
<td>N-acetylmuramoyl-L-alanin-amidase</td>
<td>13</td>
<td>59146.2</td>
<td>4.86</td>
<td>slr0891</td>
</tr>
<tr>
<td>4</td>
<td>Probable porin; major outer membrane protein</td>
<td>14</td>
<td>48666.9</td>
<td>5.19</td>
<td>slr1540</td>
</tr>
<tr>
<td>5</td>
<td>GumB protein</td>
<td>15</td>
<td>54125.4</td>
<td>4.97</td>
<td>slr1532</td>
</tr>
<tr>
<td>6</td>
<td>ATP synthase a subunit</td>
<td>16</td>
<td>53965.5</td>
<td>4.96</td>
<td>slr1326</td>
</tr>
<tr>
<td>7</td>
<td>nitrate transport 45kD protein</td>
<td>17</td>
<td>41835.1</td>
<td>4.79</td>
<td>slr1530</td>
</tr>
<tr>
<td>8</td>
<td>amino-acid ABC transporter binding protein</td>
<td>18</td>
<td>32830.1</td>
<td>4.25</td>
<td>slr2024</td>
</tr>
<tr>
<td>9</td>
<td>hypothetical protein</td>
<td>19</td>
<td>28832.3</td>
<td>4.88</td>
<td>slr1835</td>
</tr>
<tr>
<td>10</td>
<td>hypothetical protein</td>
<td>20</td>
<td>28029.6</td>
<td>4.83</td>
<td>slr1272</td>
</tr>
<tr>
<td>11</td>
<td>Probable porin; major outer membrane protein</td>
<td>21</td>
<td>27011.9</td>
<td>7.93</td>
<td>slr0431</td>
</tr>
<tr>
<td>12</td>
<td>hypothetical protein</td>
<td>22</td>
<td>27011.9</td>
<td>7.93</td>
<td>slr0431</td>
</tr>
<tr>
<td>13</td>
<td>plastoquinol-plastocyanin reductase</td>
<td>23</td>
<td>18996.5</td>
<td>4.87</td>
<td>slr1316</td>
</tr>
<tr>
<td>14</td>
<td>hypothetical protein</td>
<td>24</td>
<td>18689.9</td>
<td>8.96</td>
<td>slr0013</td>
</tr>
<tr>
<td>15</td>
<td>hypothetical protein</td>
<td>25</td>
<td>17848.8</td>
<td>6.15</td>
<td>slr1307</td>
</tr>
<tr>
<td>16</td>
<td>transcriptional repressor SmtB</td>
<td>26</td>
<td>15083.4</td>
<td>6.41</td>
<td>slr0792</td>
</tr>
<tr>
<td>17</td>
<td>carbon dioxide concentrating mechanism protein CcmK</td>
<td>27</td>
<td>12102</td>
<td>6.11</td>
<td>slr1029</td>
</tr>
<tr>
<td>p01</td>
<td>Chloroplast import-associated channel IAP75</td>
<td>6, 1</td>
<td>92307.8</td>
<td>4.29</td>
<td>slr1227</td>
</tr>
<tr>
<td>p02</td>
<td>ATP-dependent Clp protease regulatory subunit</td>
<td>1</td>
<td>91174.9</td>
<td>5.4</td>
<td>slr0020</td>
</tr>
<tr>
<td>p03</td>
<td>General secretion pathway protein D</td>
<td>2, 3, 4</td>
<td>84642.3</td>
<td>4.64</td>
<td>slr1277</td>
</tr>
<tr>
<td>p04</td>
<td>Probable porin; major outer membrane protein</td>
<td>2, 3, 4, 5</td>
<td>67601.3</td>
<td>4.47</td>
<td>slr1841</td>
</tr>
<tr>
<td>p05</td>
<td>DNA ligase</td>
<td>3, 4, 5</td>
<td>61469</td>
<td>5.07</td>
<td>slr1583</td>
</tr>
<tr>
<td>p06</td>
<td>Hypothetical protein</td>
<td>3, 4, 5</td>
<td>57735.4</td>
<td>4.81</td>
<td>slr1270</td>
</tr>
<tr>
<td>p07</td>
<td>ATP synthase b subunit</td>
<td>4, 5, 6, 7, 8</td>
<td>51733.3</td>
<td>4.89</td>
<td>slr1329</td>
</tr>
<tr>
<td>p08</td>
<td>Negative aliphatic amidase regulator</td>
<td>9, 10</td>
<td>48359.8</td>
<td>4.82</td>
<td>slr0447</td>
</tr>
<tr>
<td>p09</td>
<td>Twitching mobility protein</td>
<td>11, 12, 13</td>
<td>47899.1</td>
<td>5.84</td>
<td>slr1533</td>
</tr>
<tr>
<td>p10</td>
<td>NADH dehydrogenase subunit 7</td>
<td>14, 15, 16, 17</td>
<td>45534.7</td>
<td>5.75</td>
<td>slr0261</td>
</tr>
</tbody>
</table>
Table 1. (continued)

| p11 | Hypothetical protein peptidyl-prolyl cis-trans isomerase | 98 | 6, 7, 8, 9 | 44716.3 | 4.42 | slr0731 |
| p12 | Protein synthesis elongation factor Tu | 14, 25 | 9, 10, 11, 12, 13, 14, 15 | 43909.7 | 4.71 | slr0408 |
| p13 | putative oxalate decarboxylase | 86 | 11, 12, 13 | 43151.2 | 5.16 | slr1099 |
| p14 | putative polar amino acid transport system substrate-binding protein | 32 | 4.8 | slr1358 |
| p15 | Protease HhoA | 87 | 8, 10 | 41363.6 | 6.34 | slr1679 |
| p16 | Membrane protein | 16 | 23 | 40679.8 | 4.51 | slr1274 |
| p17 | Twitching motility protein | 17, 18 | 40616.2 | 6.57 | slr0161 |
| p18 | Phosphate-binding periplasmic protein precursor (PBP) | 19 | 21 | 40023.4 | 4.57 | slr0680 |
| p19 | Iron transport protein | 11 | 39370.4 | 4.87 | slr1295 |
| p20 | Phosphofructokinase | 12 | 38588.4 | 5.67 | slr1196 |
| p21 | Hypothetical protein | 20, 21 | 16, 17, 18, 19 | 38270.8 | 7.82 | slr1306 |
| p22 | LysR transcriptional regulator | 22, 23 | 29, 38 | 38017.7 | 5.54 | slr0998 |
| p23 | putative homolog of plant HCF136, which is essential for stability or assembly of photosystem II (ycf48) | 10 | 37291.1 | 4.76 | slr2034 |
| p24 | KETOL-ACID REDUCTOISOMERASE (ACETOHYDROXY-ACID ISOMEROXIDUCASE) (ALPHA-KETO-BETA-HYDROXYLACIL REDUCTOISOMERASE) | 81 | 35822.1 | 4.91 | slr1363 |
| p25 | MorR protein | 84 | 35091.8 | 5.06 | slr1416 |
| p26 | Hypothetical protein | 24, 25, 26, 27, 31, 32, 33, 38, 93 | 34915.2 | 4.95 | slr0151 |
| p27 | Carbonic anhydrase | 34 | 34806.6 | 5.45 | slr1915 |
| p28 | Hypothetical protein | 28, 33, 76 | 32677 | 5.38 | slr0670 |
| p29 | Phycocyanin associated linker protein | 29, 30, 31, 33, 32, 83 | 32520.8 | 9.35 | slr1580 |
| p30 | Hypothetical protein | 82 | 31790.6 | 4.83 | slr0848 |
| p31 | Carbonic anhydrase | 73 | 17 | 31204.5 | 5.13 | slr0244 |
| p32 | Hypothetical protein | 35 | 30761 | 5.8 | slr1347 |
| p33 | Phosphate transport ATP-binding protein PstB | 34, 35 | 14, 16 | 30210.5 | 5.75 | slr0683 |
Table 1. (continued)

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Accession</th>
<th>p</th>
<th>ecoli</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>p36</td>
<td>30S ribosomal protein S2</td>
<td>12, 13</td>
<td>30150.3</td>
<td>5.14</td>
<td>sll1260</td>
<td></td>
</tr>
<tr>
<td>p37</td>
<td>Hypothetical protein</td>
<td>22</td>
<td>30067.1</td>
<td>4.22</td>
<td>sll1276</td>
<td></td>
</tr>
<tr>
<td>p38</td>
<td>Hypothetical protein</td>
<td>30</td>
<td>30093.2</td>
<td>4.15</td>
<td>sll1275</td>
<td></td>
</tr>
<tr>
<td>p39</td>
<td>Photosystem II manganese-stabilizing polypeptide</td>
<td>30, 31, 15, 16</td>
<td>29911.8</td>
<td>4.82</td>
<td>sll0427</td>
<td></td>
</tr>
<tr>
<td>p40</td>
<td>Phosphate transport ATP-binding protein PsIB</td>
<td>78, 34, 79</td>
<td>29244.6</td>
<td>5.13</td>
<td>sll0684</td>
<td></td>
</tr>
<tr>
<td>p41</td>
<td>Septum site-determining protein MnD</td>
<td>18</td>
<td>29047.2</td>
<td>5.26</td>
<td>sll0289</td>
<td></td>
</tr>
<tr>
<td>p42</td>
<td>Chloroplast membrane-associated 30 kD protein</td>
<td>38, 40</td>
<td>28905.5</td>
<td>4.95</td>
<td>sll0617</td>
<td></td>
</tr>
<tr>
<td>p43</td>
<td>NADH-ubiquinone oxidoreductase subunit PsbG</td>
<td>39</td>
<td>27345.7</td>
<td>8.31</td>
<td>sll1260</td>
<td></td>
</tr>
<tr>
<td>p44</td>
<td>Hypothetical protein</td>
<td>40</td>
<td>26721.6</td>
<td>4.91</td>
<td>sll0552</td>
<td></td>
</tr>
<tr>
<td>p45</td>
<td>Elongation factor TS</td>
<td>74, 75, 19, 20, 21</td>
<td>24231</td>
<td>4.82</td>
<td>sll1261</td>
<td></td>
</tr>
<tr>
<td>p46</td>
<td>NarL subfamily</td>
<td>104, 22</td>
<td>24046.5</td>
<td>5.33</td>
<td>sll1909</td>
<td></td>
</tr>
<tr>
<td>p47</td>
<td>Methyltransferase</td>
<td>41</td>
<td>23071.7</td>
<td>6.28</td>
<td>sll0629</td>
<td></td>
</tr>
<tr>
<td>p48</td>
<td>SOS function regulatory protein</td>
<td>101, 22</td>
<td>22744.3</td>
<td>5.84</td>
<td>sll1626</td>
<td></td>
</tr>
<tr>
<td>p49</td>
<td>ATP-dependent Clp protease proteolytic subunit</td>
<td>103</td>
<td>22396.9</td>
<td>5.74</td>
<td>sll0165</td>
<td></td>
</tr>
<tr>
<td>p50</td>
<td>Light repressed protein</td>
<td>36, 37, 38</td>
<td>21895</td>
<td>6.07</td>
<td>sll0947</td>
<td></td>
</tr>
<tr>
<td>p51</td>
<td>chromosome partitioning protein, ParA family</td>
<td>45</td>
<td>21699.2</td>
<td>8.83</td>
<td>sll0929</td>
<td></td>
</tr>
<tr>
<td>p52</td>
<td>complex 23K protein PsbP homolog (psbP2)</td>
<td>99</td>
<td>20747.3</td>
<td>4.55</td>
<td>sll1418</td>
<td></td>
</tr>
<tr>
<td>p53</td>
<td>Hypothetical protein</td>
<td>46</td>
<td>20576.4</td>
<td>5.92</td>
<td>sll0362</td>
<td></td>
</tr>
<tr>
<td>p54</td>
<td>Adenylate kinase</td>
<td>25</td>
<td>20453.6</td>
<td>5.39</td>
<td>sll1815</td>
<td></td>
</tr>
<tr>
<td>p55</td>
<td>CheY subfamily</td>
<td>47, 48</td>
<td>20232.9</td>
<td>6.34</td>
<td>sll2024</td>
<td></td>
</tr>
<tr>
<td>p56</td>
<td>50S RIBOSOMAL PROTEIN L5</td>
<td>49</td>
<td>20230.5</td>
<td>9.69</td>
<td>sll1808</td>
<td></td>
</tr>
<tr>
<td>p57</td>
<td>Hypothetical protein</td>
<td>19</td>
<td>20183</td>
<td>4.52</td>
<td>sll1338</td>
<td></td>
</tr>
<tr>
<td>p58</td>
<td>ATP synthase d subunit</td>
<td>90, 44, 26, 27</td>
<td>20093.4</td>
<td>6.33</td>
<td>sll1325</td>
<td></td>
</tr>
<tr>
<td>p59</td>
<td>Pyrimidine operon regulatory protein PyrR</td>
<td>50</td>
<td>19946.4</td>
<td>9.15</td>
<td>sll0368</td>
<td></td>
</tr>
<tr>
<td>p60</td>
<td>photosystem I assembly related protein (ycfB)</td>
<td>100</td>
<td>19878.3</td>
<td>4.72</td>
<td>sll0823</td>
<td></td>
</tr>
<tr>
<td>p61</td>
<td>ATP synthase subunit b</td>
<td>51</td>
<td>19804.6</td>
<td>5.08</td>
<td>sll1324</td>
<td></td>
</tr>
<tr>
<td>p62</td>
<td>ACETOLACTATE SYNTHASE SMALL SUBUNIT (AHAS) (ACETOHYDROXY-ACID SYNTHASE SMALL SUBUNIT) (ALS)</td>
<td>102</td>
<td>18871</td>
<td>5.36</td>
<td>sll0065</td>
<td></td>
</tr>
<tr>
<td>p63</td>
<td>ribosomal protein L10</td>
<td>52</td>
<td>18675.7</td>
<td>8.64</td>
<td>sll1745</td>
<td></td>
</tr>
<tr>
<td>p64</td>
<td>ORF 157 (AA 1-157)</td>
<td>39, 40, 17</td>
<td>18288.9</td>
<td>4.65</td>
<td>sll1281</td>
<td></td>
</tr>
<tr>
<td>Page</td>
<td>Description</td>
<td>Accession</td>
<td>Molecular Weight (kDa)</td>
<td>pI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
<td>------------------------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p65</td>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p66</td>
<td>Phycocyanin beta subunit</td>
<td>53 41</td>
<td>18231.9</td>
<td>4.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p67</td>
<td>Phycocyanin b subunit</td>
<td>54, 55</td>
<td>18154.6</td>
<td>4.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p68</td>
<td>Hypothetical protein</td>
<td>57 67</td>
<td>17058.5</td>
<td>8.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p69</td>
<td>Phycocyanin a subunit</td>
<td>58, 59</td>
<td>17586.7</td>
<td>5.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p70</td>
<td>Allophyocyanin alpha subunit</td>
<td>60</td>
<td>17354.8</td>
<td>4.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p71</td>
<td>Allopbhcyanin b chain</td>
<td>61 30</td>
<td>17215.7</td>
<td>5.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p72</td>
<td>Hypothetical protein</td>
<td>59</td>
<td>16477.1</td>
<td>6.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p73</td>
<td>ATP synthase subunit b</td>
<td>62</td>
<td>16245.5</td>
<td>5.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p74</td>
<td>Hypothetical protein</td>
<td>63 64</td>
<td>15796.1</td>
<td>5.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p75</td>
<td>Photosystem I subunit II</td>
<td>64, 65,</td>
<td>15643.9</td>
<td>8.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>66, 67,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>68 69,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>70 36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p76</td>
<td>Hypothetical protein</td>
<td>110</td>
<td>15240.5</td>
<td>4.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p77</td>
<td>Hypothetical protein</td>
<td>71</td>
<td>15194.2</td>
<td>5.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p78</td>
<td>ATPase subunit epsilon</td>
<td>56, 72</td>
<td>14580.6</td>
<td>5.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p79</td>
<td>Hypothetical protein</td>
<td>66</td>
<td>14408.4</td>
<td>4.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p80</td>
<td>12 kDa extrinsic protein of photosystem II</td>
<td>63 31</td>
<td>14231.3</td>
<td>4.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p81</td>
<td>Hypothetical protein</td>
<td>45, 50,</td>
<td>14077.7</td>
<td>4.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>52, 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p82</td>
<td>Ribosomal protein L12</td>
<td>73 47</td>
<td>13259.4</td>
<td>4.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p83</td>
<td>Ribulose biphosphate carboxylase small subunit</td>
<td>74 105</td>
<td>13239.1</td>
<td>5.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p84</td>
<td>3OS ribosomal protein S6</td>
<td>73</td>
<td>13237.3</td>
<td>7.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p85</td>
<td>Photosystem II 13 kD protein</td>
<td>75, 76</td>
<td>12590.4</td>
<td>4.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p86</td>
<td>PII protein</td>
<td>77, 78,</td>
<td>12397.4</td>
<td>6.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>79 34,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>106 30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>38 35,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p87</td>
<td>Carbon dioxide concentrating mechanism protein CcmK</td>
<td>57</td>
<td>11902.7</td>
<td>5.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p88</td>
<td>Hypothetical protein</td>
<td>39</td>
<td>11450</td>
<td>5.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p89</td>
<td>Carbon dioxide concentrating mechanism protein CcmK</td>
<td>80 58</td>
<td>11134.8</td>
<td>5.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p90</td>
<td>Hypothetical protein</td>
<td>61</td>
<td>10946.7</td>
<td>4.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p91</td>
<td>Photosystem I subunit VII</td>
<td>83, 84</td>
<td>88283.3</td>
<td>5.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p92</td>
<td>Hypothetical protein</td>
<td>85</td>
<td>82886.9</td>
<td>6.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p93</td>
<td>DNA gyrase subunit A</td>
<td>86</td>
<td>8171.5</td>
<td>8.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p94</td>
<td>Photosystem I subunit IV</td>
<td>87, 88</td>
<td>8145.2</td>
<td>8.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>107 36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p95</td>
<td>Hypothetical protein</td>
<td>89</td>
<td>7220.4</td>
<td>8.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Topology prediction of identified proteins in integral fraction and hypothetical proteins from peripheral fraction

<table>
<thead>
<tr>
<th>Protein</th>
<th>Potential transmembrane segments</th>
<th>Size</th>
<th>Fraction*</th>
</tr>
</thead>
<tbody>
<tr>
<td>slr1506</td>
<td>56-76</td>
<td>1.121</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>382-402</td>
<td>1.218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>443-463</td>
<td>0.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>536-556</td>
<td>1.163</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11-31</td>
<td>1.534</td>
<td></td>
</tr>
<tr>
<td>slr1908</td>
<td>292-312</td>
<td>0.777</td>
<td>591</td>
</tr>
<tr>
<td></td>
<td>513-533</td>
<td>1.246</td>
<td></td>
</tr>
<tr>
<td>slr0224</td>
<td>4-24</td>
<td>1.690</td>
<td>298</td>
</tr>
<tr>
<td>slr1053</td>
<td>1-21</td>
<td>1.788</td>
<td>267</td>
</tr>
<tr>
<td>slr1272</td>
<td>8-28</td>
<td>1.785</td>
<td>254</td>
</tr>
<tr>
<td>slr0431</td>
<td>20-40</td>
<td>1.579</td>
<td>250</td>
</tr>
<tr>
<td>slr1338</td>
<td>5-25</td>
<td>1.738</td>
<td>167</td>
</tr>
<tr>
<td>slr1307</td>
<td>4-24</td>
<td>1.863</td>
<td>175</td>
</tr>
<tr>
<td>slr0772</td>
<td>381-401</td>
<td>0.868</td>
<td>546</td>
</tr>
<tr>
<td></td>
<td>449-469</td>
<td>0.992</td>
<td></td>
</tr>
<tr>
<td></td>
<td>472-492</td>
<td>1.079</td>
<td></td>
</tr>
<tr>
<td>slr0151</td>
<td>30-50</td>
<td>1.767</td>
<td>320</td>
</tr>
<tr>
<td>slr1270</td>
<td>12-32</td>
<td>1.957</td>
<td>526</td>
</tr>
<tr>
<td>slr1306</td>
<td>6-26</td>
<td>1.094</td>
<td>335</td>
</tr>
<tr>
<td>slr1358</td>
<td>13-33</td>
<td>2.295</td>
<td>394</td>
</tr>
<tr>
<td>slr1835</td>
<td>16-36</td>
<td>1.985</td>
<td>265</td>
</tr>
<tr>
<td>slr0670</td>
<td>23-43</td>
<td>0.679</td>
<td>291</td>
</tr>
<tr>
<td>slr0013</td>
<td>19-39</td>
<td>2.350</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>11-31</td>
<td>1.569</td>
<td></td>
</tr>
<tr>
<td>slr1450</td>
<td>61-81</td>
<td>0.636</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>119-139</td>
<td>1.151</td>
<td></td>
</tr>
<tr>
<td>slr1029</td>
<td>1-21</td>
<td>0.878</td>
<td>111</td>
</tr>
<tr>
<td>slr1316</td>
<td>18-38</td>
<td>1.612</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>99-119</td>
<td>0.735</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10-30</td>
<td>1.050</td>
<td></td>
</tr>
<tr>
<td>slr1277</td>
<td>162-182</td>
<td>0.618</td>
<td>785</td>
</tr>
<tr>
<td></td>
<td>512-532</td>
<td>1.175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6-26</td>
<td>2.052</td>
<td></td>
</tr>
<tr>
<td></td>
<td>42-62</td>
<td>1.315</td>
<td></td>
</tr>
<tr>
<td>slr0891</td>
<td>107-127</td>
<td>1.019</td>
<td>591</td>
</tr>
<tr>
<td></td>
<td>133-153</td>
<td>1.080</td>
<td></td>
</tr>
<tr>
<td></td>
<td>541-561</td>
<td>0.852</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14-34</td>
<td>1.081</td>
<td></td>
</tr>
<tr>
<td>slr1227</td>
<td>234-254</td>
<td>0.819</td>
<td>861</td>
</tr>
<tr>
<td></td>
<td>498-518</td>
<td>1.426</td>
<td></td>
</tr>
<tr>
<td></td>
<td>785-805</td>
<td>0.790</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11-31</td>
<td>1.540</td>
<td></td>
</tr>
<tr>
<td>slr1581</td>
<td>91-111</td>
<td>0.782</td>
<td>504</td>
</tr>
<tr>
<td></td>
<td>126-146</td>
<td>0.623</td>
<td></td>
</tr>
<tr>
<td></td>
<td>483-503</td>
<td>1.085</td>
<td></td>
</tr>
<tr>
<td>slr1328</td>
<td>305-325</td>
<td>0.638</td>
<td>503</td>
</tr>
<tr>
<td>slr1638</td>
<td>6-26</td>
<td>1.588</td>
<td>139</td>
</tr>
<tr>
<td>slr1862</td>
<td>119-139</td>
<td>1.890</td>
<td>143</td>
</tr>
<tr>
<td>slr0731</td>
<td>378-398</td>
<td>1.163</td>
<td>402</td>
</tr>
<tr>
<td>slr1275</td>
<td>39-59</td>
<td>1.864</td>
<td>275</td>
</tr>
<tr>
<td></td>
<td>39-59</td>
<td>2.033</td>
<td></td>
</tr>
</tbody>
</table>

* I denote integral fraction, p denote peripheral fraction
<table>
<thead>
<tr>
<th>protein</th>
<th>profile</th>
<th>region</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>slr0038</td>
<td>Mitochondrial energy transfer proteins(carrier protein)</td>
<td>9-18</td>
<td>Family</td>
</tr>
<tr>
<td>slr0848</td>
<td>ATP synthase B/B' CF (0)</td>
<td>17-71</td>
<td>Family</td>
</tr>
<tr>
<td>slr1506</td>
<td>Esterase/lipase/thioesterase</td>
<td>232-343</td>
<td>Family</td>
</tr>
<tr>
<td>slr1306</td>
<td>ATP/GTP-binding site motif A</td>
<td>9-16</td>
<td>Domain</td>
</tr>
<tr>
<td></td>
<td>polysaccharide deacytylase</td>
<td>20-165</td>
<td>Family</td>
</tr>
<tr>
<td>slr0151</td>
<td>RNA-binding region RNP-1 (RNA recognition motif)</td>
<td>165-172</td>
<td>Domain</td>
</tr>
<tr>
<td></td>
<td>TPR repeat</td>
<td>125-192</td>
<td>Repeat</td>
</tr>
<tr>
<td>slr1118</td>
<td>Protein of unknown function DUF74</td>
<td>1-108</td>
<td>Family</td>
</tr>
<tr>
<td>slr1835</td>
<td>Curli production assembly/transport component CsgG</td>
<td>2-242</td>
<td>Family</td>
</tr>
<tr>
<td>slr0244</td>
<td>Usp domain</td>
<td>1-132</td>
<td>Domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>139-280</td>
<td></td>
</tr>
<tr>
<td>slr0362</td>
<td>HAD-superfamily hydrolase, subfamily IIIA</td>
<td>41-152</td>
<td>Family</td>
</tr>
<tr>
<td>slr0670</td>
<td>Usp domain</td>
<td>1-134</td>
<td>Domain</td>
</tr>
<tr>
<td>slr1034</td>
<td>Single-strand binding protein</td>
<td>1-100</td>
<td>Family</td>
</tr>
<tr>
<td>slr1270</td>
<td>Outer membrane efflux protein</td>
<td>103-318</td>
<td>Family</td>
</tr>
<tr>
<td></td>
<td></td>
<td>330-525</td>
<td></td>
</tr>
<tr>
<td>slr1276</td>
<td>Proline-rich region</td>
<td>237-275</td>
<td>Family</td>
</tr>
</tbody>
</table>
Table 4. Summary of protein identification

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Number of</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>spots analyzed</td>
<td>Spots identified</td>
<td>proteins identified</td>
<td>proteins without known function</td>
<td>proteins with isoforms</td>
</tr>
<tr>
<td>Peripheral</td>
<td>310</td>
<td>249</td>
<td>95</td>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td>Integral</td>
<td>132</td>
<td>63</td>
<td>29</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Sum*</td>
<td>442</td>
<td>312</td>
<td>112</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>Previously Indetified#</td>
<td>116</td>
<td>78</td>
<td>51</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Total*</td>
<td>558</td>
<td>390</td>
<td>128</td>
<td>38</td>
<td>47</td>
</tr>
</tbody>
</table>

* The number is the sum of all the numbers in each fraction minus the overlapped number of proteins.

# See reference
Figure Legends

Figure 1. Separation of peripheral and integral thylakoid proteins. T is the thylakoid membrane containing 10 μg chlorophyll. P is the peripheral protein extracted with 8 M urea, precipitated with 10% TCA and solubilized with multi-surfactant solution. I is the integral fraction washed with acetone and followed by solubilization with multi-surfactant solution. The amount of samples loaded for P and I fractions came from same amount of cells. All the samples were incubated with denaturing dye (2% SDS and 0.1 M DTT) at room temperature for 4 h.

Figure 2. Separation of the peripheral thylakoid membrane proteins by 2-DE. The first dimension included 18 cm IPG strip with pH range (a) 3-10, nonlinear, (b) 4-7, (c) 4-5, (d) 5-6. The second dimension included 12-18% gradient SDS-PAGE gel for all. All labeled protein spots in the 2-D gel were analyzed with peptide mass fingerprinting.

Figure 3. Separation of the integral thylakoid membrane proteins by 2-DE. The first dimension included 18 cm IPG strip with pH range (a) 3-10, nonlinear, (b) 4-7. The second dimension included 12-18% gradient SDS-PAGE gel for all. All labeled protein spots in the 2-D gel were analyzed with peptide mass fingerprinting.

Figure 4. The distribution of GRAVY among (a) identified peripheral thylakoid membrane proteins, (b) identified integral thylakoid membrane proteins, (c) proteins from the whole proteome of *Synechocystis* sp. PCC 6803. The GRAVY value and the occurrence were calculated with a computer program written in C++.

Figure 5. The relationship between GRAVY and the number of TM. The number of TM for the proteins from the whole proteome of *Synechocystis* were predicted using a automated TMpred program.
Fig. 1
Fig. 2
Fig. 3
Fig. 5
CHAPTER 5. FUNCTIONAL STUDY OF HYPOTHETICAL THYLAKOID PROTEINS FROM SYNECHOCYSTIS SP. PCC 6803

A paper to be submitted to the Journal of Biological Chemistry
Yingchun Wang*, Gaozhong Shen§, Donald A. Bryant§, Parag R. Chitnis*²

SUMMARY
The thylakoid membranes of Synechocystis sp. PCC 6803 are estimated to contain about 600 proteins, most of which are hypothetical (1). Our previous proteomic analysis has identified 128 thylakoid proteins, 38 of which have no known functions. Knock-out mutants of ten open reading frames were obtained. The preliminary screening results indicate that only one mutant (H1) which has a deletion on the ORF slr0110 has a conditional growth deficiency phenotype. The H1 mutant is sensitive to glucose and light; its function is assumed to be related with photosynthesis or respiration. The H1 mutant can be rescued by the adding of DCMU or rotenone to the growth medium. Therefore, we postulate that the gene plays roles for the maintenance of the poised redox state of plastoquinone (PQ)¹ pool. The total amount of PSI and PSII in the stressed H1 strain decreased in the mutant thylakoid membranes, but the specific activity of PSII was not decreased. Chlorophyll fluorescence induction analysis indicates that the PQ pool of H1 is more reduced than that of wild type strain, the reduction is due to electrons from PSII or NADH dehydrogenase. The deletion of slr0110 changed the transcription level of downstream genes, resulted in an increase in mRNA, but the overall transcription level of the whole genome decreased compared with wild type strain. The proteome analysis indicates that the amount of two PSII subunits Cyt c550 and the 12 kD extrinsic PS II protein increased in H1 strain, which could lead to the

¹ This work is supported in part by grants from Iowa Corn Promotion Board and from the Vice-Provost for Research at ISU. Journal Paper No. J-XXXXX of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3416 and supported in part by Hatch Act and State of Iowa funds.
² *Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University; §Department of Biochemistry and Molecular Biology, Pennsylvania State University
³ Abbreviations: Cyt c550, cytochrome c550; Cyt bf, cytochrome bf; Fv, variable fluorescence; Fm, Maximal fluorescence; IPG, Immobilized pH gradient; PQ, plastoquinone.
activity increase of PSII in H1. The NAD(P)H-quinone oxidoreductase subunit K homolog 2 (encoded by plasmid gene psbG2) which was found to be expressed only in the photomixotrophic sensitive psbG1 deletion mutant, was also found to be specifically expressed in H1 strain. Therefore, it is highly possible the function of the hypothetical proteins encoded by the ORF slr0110 is related with psbG1 gene.

INTRODUCTION

Cyanobacterium *Synechocystis* sp. PCC6803 has been used as a model system to conduct proteomic studies of membrane proteins (2-4). The unicellular organism has been widely used for molecular genetic and biochemical studies of photosynthesis and respiration. It contains a large fraction of photosynthetic thylakoid membranes and its 3.57 Mb genome has been completely sequenced (1). The cyanobacterium can easily take up foreign DNA and insert it to its genome by homologous recombination, making it a powerful model system for a functional genomic approach. Therefore, the thylakoids of *Synechocystis* provide an ideal system for the global studies of membrane proteins.

About 1000 ORFs of *Synechocystis* have been estimated to encode membrane proteins (5,6). The membrane system of *Synechocystis* sp. PCC 6803 is similar to that of chloroplasts in higher plants, it contains both cytoplasmic and thylakoid membrane. The cytoplasmic membrane forms the inner boundary of the periplasmic space and is estimated to contain about 400 different proteins, some of which are involved in respiratory pathway. The thylakoid membranes contains the largest fraction of the membrane system and are estimated to contain 600 different proteins involved in both photosynthetic and respiratory pathways as well as many other biological processes. Proteins involved in the respiratory electron transport chain typically includes NADPH-preferring type I dehydrogenase (NDH-1), succinate dehydrogenase (SDH) and cytochrome aa3-type terminal oxidase are typically included in the respiratory electron transport chain, whereas photosynthetic electron transport chain typically includes photosystem I (PS I) and photosystem II (PS II) protein complexes. The two electron transport chains have been found interwoven with each other and share the plastoquinone (PQ) pool, the Cyt b6f and the soluble electron carriers. The PQ pool receives electrons from NADH dehydrogenase on the respiratory chain or from PSII on the photosynthetic chain. The redox state of PQ pool is determined by the rate of electrons
flowing in and out of it. The poised redox state of the PQ pool is crucial for the growth and the state transition of Synechocystis (7). An extremely reduced or oxidized PQ pool may affect the growth of the Synechocystis cells, which can reasonably explain why PSI less mutants are light sensitive as electron transport into the PQ pool from PSII cannot be pulled out by less PSI (8).

Electron transport in and out of the PQ pool is very complex and many pathways exist. Three interwoven pathways have been considered to be the dominant pathways in Synechocystis. The three pathways are: (1) The linear photosynthetic pathway in which electrons are transported into PQ pool from PSII and then finally pulled out of PQ pool by PSI; (2) The respiratory pathway in which electrons are transported into PQ pool from NADPH dehydrogenase and succinate dehydrogenase and finally pulled out by terminal oxidase; (3) The cyclic pathway in which electrons transport around PSI. The electrons generated from one pathway can easily transport to the other at the shared points. For instance, the electrons generated from PSII can feed to cytochrome oxidase in the PS I-less mutant, and the electrons generated from the respiratory pathway can reduce the acceptor side of PSII if the terminal oxidase is blocked by KCN.

Although the thylakoid proteins play important roles in photosynthesis, respiration and other metabolic pathways, most of them are still unknown for their functions. To investigate the functions of these unknown proteins, we performed proteomic study as well as mutagenetic analysis. The thylakoid membrane has been separated into integral and peripheral fractions, proteins from both fractions have been resolved into 2-D gels and identified by MALDI-TOF mass spectrometry coupled with database search. Thirty eight hypothetical proteins that are encoded by putative ORFs have been identified. Twelve hypothetical proteins that have no existing mutants referenced in Cyanobase were chosen for mutagenetic analysis, the corresponding ORFs were inactivated through knockout mutation and 10 segregated mutants were obtained. The growth analysis showed nine of the mutants have no obvious phenotype under photoautotrophic and photomixotrophic conditions. Analysis on pigment contents also did not show much difference with wild type. Only one mutant, which has the inactivation of the ORF slr0110, showed obvious phenotype on both growth and pigment contents. The detailed characterization indicated that this mutant is
sensitive to glucose and light. We assume that the function of the gene slr0110 is involved in electron transport in both photosynthesis and respiration and crucial for the poised redox state of PQ pool.

EXPERIMENTAL PROCEDURES

Deletion of slr0110 Genomic DNA from *Synechocystis* sp. PCC 6803 cells was isolated by a previously described method (9). A pair of primers were synthesized to amplify a 1031 bp DNA fragment containing upstream and a large fraction of the ORF slr0110 by polymerase chain reaction (PCR) (5'-GTTGCGGTCAACGTTAATCC-3', 5'-CCGTCGATAATCCGATTGG -3'). The amplified DNA fragment was purified from agarose gel prior to be inserted into pGEM-T vector (Promega) with blunt end ligation. The resulting construct was digested with Hpa I that has two closely located sites in the coding region of slr0110. Finally, a 2.0 kb spectinomycin resistant gene cassette was inserted into the DNA fragment to replace the sequence between the two HpaI sites through blunt end ligation. The final construct containing the interrupted slr0110 was transformed to *Synechocystis* cells using a previously described method (10). After several generations of segregation on BG-11 medium plates with 50 μg/ml spectinomycin under low light intensity, stable mutant strains were generated. Genomic DNA was extracted from the mutant strain; inactivation of slr0110 was confirmed using PCR analysis.

Growth and rescue analysis Wild type and the H1 strain of *Synechocystis* sp. PCC6803 cells were grown in liquid BG11 containing spectinomycin (50 μg/ml) at 30 °C, with a final concentration of 50 mM glucose for photomixotrophic or no glucose for photoautotrophic growth. The cells were grown under high light intensity (>120 μ mol. m⁻².s⁻¹), medium light intensity (40-50 μ mol. m⁻².s⁻¹) and low light intensity (<10 μ mol. m⁻².s⁻¹). The growth began from the start point of OD₇₅₀=0.05, the concentration of cells was measured by reading the value of OD₇₅₀ using spectrophotometer every 2 4 hours. DCMU was used to rescue H1 strain grown under medium light or high light intensity with a final concentration of 50 μM in liquid BG11. The Plant toxin rotenone was used to rescue the H1 strain stressed with glucose with a final concentration of 50 μM in liquid BG 11. Chlorophyll a and carotene concentrations were estimated according to previously published methods (11,12).
Reverse transcription-polymerase chain reaction (RT-PCR)  Total RNA was prepared from exponentially growing cells according to a published method (13). Total RNA (5 μg) was dissolved in the solution containing 40 mM Tris, 6 mM MgCl₂, 10 mM NaCl, 10 mM CaCl₂ and 0.1% diethylpyrocarbonate prior to be treated with RNase-free DNase at 37 °C for 1 h. After that, the total RNA was further purified as previously described (14). The reverse transcription reaction was carried out using ThermoScript™ RT-PCR System as described by the manufacturer (Invitrogen). Random hexamer were used as primers. Primers that were specific to the upstream and downstream ORFs as well as slr0110 itself were synthesized for PCR (Table 1), each pair of primers will be able to amplify a large fraction of the specific ORF sequence.

PSI and PSII activity in the whole cell  PSI activity was monitored by measuring oxygen uptake using a Hanstech oxygen electrode. Cells were suspended in 40 mM Hepes pH7.0 solution at a chlorophyll concentration of 5 μg/ml, and then aliquoted into 1 ml and shielded from light. DCMU (50 μm) and KCN (2mM) were used to block the electron flow from PSII to PSI and respiration electron transport respectively. DAD (1 mM) were used as electron donor and methyl viologen was used as electron acceptor (2mM), Oxygen uptake activity was measured at a light intensity of 3000 μmol. m⁻².s⁻² at 25 °C. The PSII activity was determined by measuring oxygen evolution in the similar way as measuring oxygen uptake. Benzoquinone (1 mM) was used as electron acceptor and PSII itself was used as electron donor.

77 K fluorescence emission spectra  The low temperature fluorescence emission spectra were measured using a SLM 8000C spectrofluorometer as described (15). Cells were harvested at the exponential phase of growth by pelleting and resuspended in a 25 mM HEPES/NaOH, pH 7.0 buffer. Cells (5 μg Chl) were diluted in 25 mM Hepes/NaOH, pH 7.0 to a final volume of 30 μl and adapted in dark for 30 minutes on ice. To the solution, 70 μl of neat glycerol was added, mixed, prior to quickly freezing in liquid nitrogen. The excitation wavelength was 435 nm. The excitation slit width was set at 4 nm and the emission slit width was set at 2 nm. The emission was scanned from 600 nm to 800 nm twice and averaged.

Chlorophyll fluorescence measurement  Chlorophyll fluorescence was measured using a PAM Chl fluorometer (Walz, Effeltrich, Germany) as described (8). The signal was
recorded and analyzed with computer using the software DA 100. The sample contains 1 O.D. intact cells resuspended in Hepes, pH 7.0 buffer. White actinic light was obtained from a Fiber-Lite source and controlled by a shutter manually.

**Isolation of total thylakoid proteins and separation by 2-D electrophoresis** The wild type and the mutant cells were grown under medium light without glucose to exponential phase and then treated with 5 mM glucose for two days. Thylakoid membranes from both strains were prepared using differential centrifugation as previously described (16). Membranes from same amount of cells were precipitated by 10% TCA in ice-cold acetone, lipids and pigments were removed though washing with 100% ice-cold acetone. The delipidated membrane were dried with vacuum and then extracted with multiple surfactant rehydration buffer (5M urea, 2M thiourea, 2mM TBP, 2% Sulfobetain 3-10, 0.5% Carrier ampholytes, 40 mM Tris, 0.001% Orange G dye), the undissolved fraction was removed by centrifugation with 76000 x g. The proteins were first separated by iso-electrical focusing (IEF) using immobilized pH strips with a pH range 4-7 (Bio-Rad). The active in-gel rehydration and focusing procedure were programmed as : 0V for 2hours, 50 V for 10 hours, 100 V for 0.5 hour, 500 V for 0.5 hour, 1000 V for 0.5 hour, 2500 V for 0.5 hour, 5000 V for 0.5 hour and the continue to focus at 8000 V until 80,000 Vh was reached. The IPG strips were equilibrated for 20 minutes prior to be transferred to the second dimension using a solution containing 6 M urea, 2% SDS, 20% glycerol, 5mM TBP, 2.5 % acrylamide monomer, 0.3 M Tris.cl (pH 8.8). The second-dimensional electrophoresis was performed using 10-18% gradient SDS-PAGE gel. Upon electrophoresis, the protein spots were stained with colloidal Coomassie blue for 48 hours, and the staining intensity was enhanced by washing with 1% acetic acid. The gels were scanned with Densitometer GS-800 (Bio-RAD) to obtain images for analysis with PDQuest software.

**Mass spectrometry and protein identification** Protein spots from stained gels were excised and subjected in-gel digestion with trypsin as we described before (4). The resulting fragments were analyzed using MALDI-TOF mass spectrometry and the obtained peptide mass fingerprint (PMF) data was used to search database for identification (4).

**RESULTS**
Inactivation of the ORF slr0110  The construct containing the ORF slr0110 interrupted by spectinomycin resistance gene cassette as diagramed in Fig. 1a was transformed to wild type *Synechocystis* cells. The completely segregated mutant was confirmed by PCR (Fig. 1b). Fig. 1b shows that a 1031 bp DNA fragment was amplified from wild type genome and a 3.3 kb fragment was amplified from the mutant genome. The increased size of PCR product from mutant genome accounts for the insertion of the 2.3 kb spectinomycin resistance gene cassette. The PCR result indicates that the ORF slr0110 has been completely interrupted by spectinomycin resistance gene cassette in the mutant strain, which was thereafter named as H1.

Growth of the H1 strain  The H1 strain was firstly analyzed for its photoautotrophic and photomixotrophic growth. The growth curves of H1 strain in different conditions are presented in Fig. 2. As shown in the Fig. 2 (a) and (b), H1 can grow in liquid BG11 in the absence of glucose under low light or medium light intensity, there is no obvious difference for the growth curve and the color of the liquid culture compared with the wild type (Fig. 3a). Analysis for pigment contents also did not show much difference. If a final concentration of 2.5 mM glucose is added to the culture under medium light intensity, the growth of the mutant will be hampered (Fig. 2 (b)). Initially, the color of the liquid culture became pale (Fig 3. b), the concentration of cells increased to 0.5 (OD$_{730}$) after 48 hours, and then decreased and finally the color of the culture become grey and the cells died. The mutant can survive under low light intensity in the presence of glucose even though the color of the cells became pale, and the concentration of mutant cells increased at the same rate as the wild type (Fig. 2. (a)). Under high light intensity, the mutant will die either in the presence of glucose or in the absence of glucose, the color of the liquid is grey, but the value of OD$_{730}$ of the H1 mutant continue to increase to 1.5, and then drop down . These results indicate that H1 mutant strain is not only sensitive to high light, but also sensitive to glucose. Medium light is not lethal to H1 mutant, but if combined with the effect of glucose, it becomes lethal. Low light intensity in conjunction with glucose has adverse effect on growth of H1 mutant to some degree, but is not lethal.

Glucose and light have a same physiological effect on *Synechocystis*, that is, both can help the cells to produce electrons. Light energy can be absorbed by PSII and generate
electrons from water, glucose can be metabolized through TCA cycle and finally generate electrons through NAD(P)H dehydrogenase or succinate dehydrogenase (17-19). Electrons from both pathways can transport into PQ pool, which we know that its redox state is very crucial for cell growth. Therefore, we postulate that too many electrons generated from light or glucose is the real factor to kill the mutant. If we block the electron transport to PQ pool from either pathway, we should be able to rescue the mutants.

The herbicide DCMU has been widely used in photosynthesis research because it can block the electron transport from PSII to the PQ pool. Our results show that DCMU can rescue the mutant cells grown under medium light and high light intensity in the presence of glucose (Fig. 2d). From the growth curve we can see that there is no difference between the mutant and the wild type in the presence of DCMU under both light intensities, even though the cells grow more slowly than the cells grown in the absence of DCMU. The pigment analysis also indicates that there was no difference between wild type and the mutant. The color of the liquid culture of H1 was very similar to that of wild type grown with or without DCMU, but very different with that of H1 strains grown with glucose and without DCMU. These results suggest that H1 mutant can be rescued by DCMU if grown under medium and high light intensity in the presence of glucose.

The other possible way to rescue H strain grows is to block electrons originated from glucose. NADH dehydrogenase is one of the proteins on respiratory pathway that can transport electrons to the PQ pool; it can be inhibited by the plant toxin rotenone, and thereby can block the electrons transport into PQ pool. The rescue experiment shows that the H1 mutant grown under medium light intensity with glucose can be rescued by 50 μM rotenone. The color of the liquid culture with rotenone is green while color of the culture without rotenone is grey-yellow (Fig. 3C), and the cells also grow faster in the presence of rotenone. Interestingly, rotenone has no obvious effect on the growth of the wild type cells, which may indicate that the wild type have other pathways to use glucose.

All the results from the growth and rescue experiments suggest that the amount of electrons generated from PSII respiratory pathway is vital for the survival of H1 mutant. We proposed that inactivation of the gene slr0110 will change the redox state of PQ pool, which eventually make the cell sensitive to light and glucose somehow.
**PSI and PSII activity** The activities of PSI and PSII are two important factors that affect the redox state of PQ pool. Electrons generated from PS II will flow into PQ pool and make it more reduced. Electrons can also be pulled out of PQ pool by PS I and make it more oxidized. To measure the PSI activity, we used the cells grown under medium light intensity without glucose and harvested at exponential phase (OD=0.6). The cells for measuring PS II activity were further stressed with 5mM glucose for 16 hours. The activity was measured with per chlorophyll for PSI and per cell for PSII. The oxygen uptake curve shows that PSI activity of H1 strain is the same as wild type (Fig. 4a), which suggested that inactivation of the ORF slr0110 did not affect the activity of PSI. Therefore, for H1 strain, the ability of pulling electrons from PQ pool by PSI is not affected, the redox state changes of PQ pool may be caused by factors other than PSI. The oxygen evolution curve shows that the activity of mutant PSII is a little bit higher than that of wild type PSII, which suggests that mutant PSII can donate electrons to PQ pool with at least the same rate as wild type PSII does if not higher. Considering the measurement was based on per cell and the chlorophyll content in H1 strain is much lower than that of wild type, the activity of PSII of H1 strain should be higher than that of wild type based on equal amount of chlorophyll. Equal activities of PSI and PSII for wild type and the mutant make their PQ pools the same redox state, which can reasonably explain why the growth curve of H1 mutant is similar to the wild type under medium light without glucose.

**77 K fluorescence emission spectra** The ratio of PSI to PSII can be determined from 77 K fluorescence emission spectra. For the cells grown under medium light without glucose, the 725 nm emission peak arising from PSI and the peaks arising from PSII at 685 and 695 are same for wild type and the mutant (Fig. 5a) (8). To examine the effect of glucose and light exerted on the PSI and PSII ratio, we grew cells under medium light without glucose to exponential phase; divide the cultures into three groups. In group 1, cells were grown under medium light for 16 hours without glucose. In group 2, cells were grown under medium light for 16 hours with glucose. In group 3, cells were grown under high light with glucose for 16 hours. Same amount of cells from each group were used to detect the 77 k fluorescence emission spectra. In group one, the 725, 695 and 685 nM peaks of the mutant are almost the same as those of wild type, this result is in agreement with the spectra obtained.
based on same amount of chlorophyll (data not shown). In group 2, the color of the cells become pale, the chlorophyll content decreased a little bit compared with that of wild type. The three maximum peaks of 77K fluorescence emission spectrum decreased a little bit for the H1 mutant, which indicates that the total amount of PSI and PSII decreased for the cells treated with glucose. In group 3, the color of cells become pale, the chlorophyll content of the mutant is just the half amount of that of wild type. The three peaks of 77K fluorescence emission spectrum of H1 mutant decreased dramatically to the half of those of wild type. This result indicates that the total amount of PSI and PSII of H1 mutant in this treatment is just equal to the half of that of wild type. The results strongly suggest that both glucose and light will affect the synthesis of chlorophylls in H1 mutant, and eventually affect the accumulation of photosynthetic complexes. High light has stronger effect than glucose on the accumulation of PSI and PSII complexes.

**Fluorescence Induction**

The rate of electrons leaving the PSII complex can be determined by monitoring variable fluorescence upon actinic illumination. \( Q_A \) is the first electron-accepting PQ of PSII; it is a quencher of fluorescence in oxidized form while it is not in reduced form (8). Changes in fluorescence yield reflect the redox state of \( Q_A \) in the time scale of milliseconds to a few seconds (8,20). \( Q_A \) accepts electrons from the reducing site of PSII, and then transfer the electrons to the other two-electron PQ acceptor \( Q_B \). Double reduction release \( Q_B \) from its binding site on PSII and diffuse to PQ pool, new oxidized PQ molecule will occupy the empty binding site and accept new electrons from \( Q_A \) again. Therefore, the redox state of \( Q_A \), in turn, depends on the availability of PQ molecules, and fluorescence induction measurement can also be used to determine the redox state of PQ pool.

Cells for fluorescence induction analysis were grown and treated in the same way as the cells for 77 K fluorescence emission spectra analysis. Fluorescence induction curves of whole cells from wild type and H1 mutant strain grown under different conditions are presented in Fig.6. In each curve, the fast phase of fluorescence rise concomitant with the shutter opening is taken as \( F_0 \). The \( F_v \) comes from the reduction of \( Q_A \). The fluorescence yield kept increasing until got saturated, when equilibrium of the electrons accepted by \( Q_A \) and the electrons acquired by PQ pool is reached. Therefore, \( F_v \) can be considered to be the volume of reduced PQ pool.
In the cells grown under medium light without glucose, Fv of H1 strain is larger than that of wild-type, whereas the FM is same. Therefore, the PQ pool of H1 strain is more reduced than that of wild-type (Fig. 6a, 6b). If the cells stressed with 5 mM glucose, the curve of fluorescence induction would be very similar to that of grown under normal conditions (Fig. 6c, 6d). In the cells stressed with high light and glucose, the Fv in both wild-type and H1 are increased, but Fv of H1 strain is still larger than that of wild-type strain (Fig. 6e, 6f). The Fm of both strains in this treatment is smaller compared with the other two groups. These results imply that under all conditions, the PQ pool of H1 strain is more reduced than that of wild-type, and the reduction increased with light intensity.

**Transcriptional analysis**  The ORF slr0110 is flanked by seven other ORFs in the genomic sequence, with the ORFs slr0111, slr0112, slr0114, slr0115 in downstream and sll0094, sll0095 in upstream (Fig. 7b). The close localization of the ORFs indicates that they may be in the same operon. Deletion of the ORF slr0110 could delete the promoter sequence or other regulating or controlling sequence for the neighboring ORFs and thereby lead to inactivation of these ORFs. If this really happens, we will be unable to tell whether the phenotype of the H1 mutant resulted from the inactivation of ORF slr0110 or from the other ORFs. Therefore, it is necessary to investigate whether the expression of these neighboring ORFs were affected or not. The expression was detected at transcription level using RT-PCR. Total RNA were prepared from same amount of wild type and H1 mutant cells grown under medium light without glucose. The quality and quantity of RNA were checked using agarose gel electrophoresis by loading RNA from the same amount of cells. As shown in Fig. 7a, the RNA bands in the lanes WT-A and WT-B are wild type RNA from two independent preparations, The RNA bands in the lanes ½ WT-A and ½ WT-B are the wild type RNA from the same preparations for the RAN in lanes WT-A and WT-B respectively, but the amount loaded was the half to the RAN in the corresponding lanes. Lane H1-A and lane H1-B, lane ½ H1-A and ½ H1-B are RNA from H1 strain; they were loaded in the same way as wild type RNA. The image analysis showed that the concentration of RNA from two parallel preparations were almost the same for both wild type and H1 strain (The amount of RNA in lane WT-A=WT-B, ¼ WT-A=1/2 WT-B, H1-A=H1-B, ½ WT-A=1/2 H1-B). Therefore, the amount variation came from manipulations during preparation is very small.
This makes it possible for us to compare the amount of RNA from H1 and wild type strains. The quantity analysis indicated that the concentration of H1 RNA is only half to that of wild type RNA based on equal amount of cells, as compared by the band intensity in lanes WT-A and H1-A or $\frac{1}{2}$ WT-A and $\frac{1}{2}$ H1-A. Therefore, it is obvious that the overall transcription level of H1 is only half to that of wild type.

The RT-PCR was performed using RNA from same amount of cells, thus we can find out the real expression level of those neighboring genes in each single cell. Five ORFs slr0111, slr0112, slr0115 and slr0094, slr0095 that locate at the downstream or upstream of the ORF slr0110 respectively were investigated for transcriptional variation. Three other ORFs sir1841, sll1703 and sll0337 that have no known contribution to the phenotype of H1 mutant and are far away from slr0110 in the genome were also investigated for transcriptional variation as internal standards. The DNA products of RT-PCR were checked with agarose gel electrophoresis as shown in Fig. 7c. For the ORF slr0110, a 500 bp DNA band was found as the product of RT-PCR using wild type RNA as template (lane1), and no band was found at the corresponding place in the lane for the product of RT-PCR using H1 RNA as template. The results indicate that the ORF slr0110 is expressed in wild type, but completely inactivated in H1 strain. Very interestingly, the three downstream ORFs slr0111, slr0112 and slr0115 are all expressed in both wild type and H1 strain, and the quantity analysis indicated that the amount of DNA products were highly increased for all of the three ORFs in H1 strain (lane WT/slr0111 and H1/slr0111, WT/slr0112 and H1/slr0112, WT/slr0115 and H1/slr0115). Since the DNA products of the three internal standards were not increased in H1 strain (The lanes for ORF sir1841, sll1703 and sll0337), it is obvious that the increase of DNA amount is completely due to the increase of the expression level of the three genes in H1 strain. Considering the overall transcription level in H1 strain is decreased, the increase of transcription level of the three genes is significant. We have no idea why the expression of these genes increased in the H1 strain and what physiological roles they may play, but believe that increased expression level of these genes will affect the growth of the mutant. There were no DNA products of RT-PCR detected for the upstream ORFs slr0094 and slr0095 from both wild type and the H1 strains; it is possible that these ORFs were not
expressed under the specific growth conditions or the expression is below the detection limit of RT-PCR.

**Proteomic analysis of wild type and H1 thylakoid proteins** A well resolved 2-D pattern was obtained for both wild type and H1 mutant thylakoid proteins respectively, about 200 spots were detected in 2-D gels stained with colloidal Coomassie blue using PDQuest analysis (Fig. 8). Some high abundant protein spots presented on both gels were chosen as landmarks such as spots w07, w08, w13, w14 and w15. Both landmark spots and differentially displayed protein spots were chosen for identification. Totally 16 spots were analyzed and identified from the 2-D gel for wild type while 4 out of 6 analyzed spots from H1 strain were identified (table 2). The spots h03, h04 were failed to be identified due to the reason as we discussed before (4).

Of the four identified spots in the h1 gel, the intensity of h05 and h06 are highly increased compared with the corresponding spots on the wild type gel. The spot h05 was identified as cytochrome c550 (Cyt c550) which was encoded by pub, h06 is the 12 KD extrinsic protein of photosystem II encoded by pubs, both proteins are the photosystem II donor-site subunits (21). Functional study indicated that Cyt C550 plays a substantial role in maintaining the stability and function of the manganese cluster, whereas the 12-kDa protein plays primarily a regulatory role in maintaining normal S-state transitions (22-24). The other extrinsic protein of PSII: the manganese-stabilizing protein (MSP) was identified in wt gel (w07) and its corresponding spot was also found in HI gel. The amount of MSP did not show much difference on both gels. All of the three proteins MSP, Cyt c550 and the 12 kDa proteins are reported to be able to enhance the thermal stability of the oxygen-evolving machinery.

The spots h01 and h02 were only found in HI gel. The spot h02 is a hypothetical protein, whereas the h01 is the product of psbG2 gene, which encodes NAD(P)H-quinone oxidoreductase subunit K homolog 2 (25). The psbG2 is a plasmid gene; its homolog psbG1 gene is located on chromosome. The psbG2 is not transcribed in wild type cells, but transcribed in psbG1 deletion mutant (26). This finding support our result that psbG2 product was found in the wild type gel. Very interestingly, the psbG1 deletion mutant not only has the expressed psbG2 product, but has very similar growth phenotype of the H1 mutant. Both
mutants can grow under autotrophic conditions, while their growth is impeded in mixotrophic medium (26). Therefore, it will be very interesting to investigate whether there any correlation between the deletion of slr0110 and the expression of psbG1.

Discussion

The function of hypothetical proteins that has no obvious phenotype in knockout mutants  In our proteomic study, 38 hypothetical thylakoid proteins have been identified. We selected 12 hypothetical proteins that have no referred mutants in Cyanobase for mutagenetic analysis. Ten mutants have been created in the similar way as described for H1 strain (table 2). The other two hypothetical proteins (ssl1690, ssl1306) that failed to be deleted from genome may be indispensable for the growth of Synechocystis. The mutants have been grown under photoautotrophic or photomixotrophic condition, the doubling time and pigment content of each mutant did not show much difference with wild type except H1 (Table 3). It seems that deletion of these genes did not produce any detrimental effects on Synechocystis cells. Therefore, these genes are “redundant” in some sense. The redundancy could be caused by at least two factors: One is that the deleted gene has one ore more homologs in the same genome, the homologs are not expressed in wild type cells but waked up by the deletion of the gene. The function of the deleted genes can be substituted by the activated homologs. Therefore, the roles of the deleted gene can be replaced by its homologous. The other factor could be the metabolic network. Many genes do not function independently, but function as a unit in a metabolic pathway or network. Dysfunction of the gene may cause one pathway or a branch in a network blocked, but other pathway or branch may be activated. Therefore, the function of this pathway may be replaced by other pathway or network.

What can the results of 77K and fluorescence tell us?  The results of 77 K fluorescence emission spectra indicate that the total amount of PSI and PSII decrease if the H1 strain is stressed with glucose and light. The amount decrease was not accompanied by the decrease of PSII activity, as shown by the fluorescence induction results. PSII from same amount of stressed H1 can provide electrons to the PQ pool as fast as stressed wild type did, if not much faster. The Oxygen evolution experiments also indicate that PSII of H1 strain has the same activity as that of wild type based on equal amount of cells. Since in each cells the
total amount of PSII of H1 stain is less than that of wild type, it is obvious that the activity of each individual PSII complex are much higher than that of wild type. This could be a good explanation to the question why the PQ pool is more reduced.

**The homologous ORF slr0229** The homology analysis using BLAST program shows that the ORF slr0110 has a homologous ORF slr0229, as shown in the sequence alignment of these two ORFs (Fig. 9). Both of the two ORFs have an ATP-GTP binding domain at the N-terminal. And these domains are highly conserved among all of the chromosome partition protein ParA family. Initially, we thought that the gene slr0110 may have a function related with chromosome partition, and the H1 strain is sensitive to some stress condition like glucose and light intensity because the chromosome can not separate and replicate. To test this hypothesis, we set up a series of stress condition for the growth of the mutant. The stress conditions include less sulfur, less nitrogen, less cooper, less iron for the growth medium. However, the stresses did not cause any obvious growth phenotype (Data not shown). Therefore, we discard this hypothesis and formulate the other one. The knockout mutant of slr0229 has also been obtained (H11 in table 3). The growth experiments showed that the mutant strain can grow photoautotrophically or photomixotrophically. The effort of trying to create slr0110, slr0229 double deletion mutant was suffered from being not successful. Therefore, it is possible that the function of slr0229 can be at least partly replaced by the gene slr0110.

**The amount variation of cyt550 and the 12 kD extrinsic PSII proteins** Deletion of slr0110 caused the increase of expression of some genes, such as the gene encodes cyt550 and the one encodes the 12 KD extrinsic PSII proteins. Both of the two proteins are related with PSII. Although we do not know what is the mechanism underlined that the amount change of these two proteins affect the activity of PSII, we know that the activity of PSII in H1 strain is increased. This result is in agreement with the other physiological analysis. We can assume that the deletion of slr0110 lead to highly expression of cyt550 and the 12 kD extrinsic PSII protein, the amount increase of the two proteins lead to the activity increase of PSII, which in turn can generate and transport more electrons to the PQ pool. The reduced PQ pool may be the primary reason that leads the H1 strain sensitive to glucose and light.
Why PQ pool of H1 is more reduced? The poised redox state PQ pool is important for balancing respiratory and photosynthetic electron transport and for regulation of the different enzymes (7). The two electron transport chains on the thylakoid membrane and their relationship to the PQ pool were diagramed in Fig. 10. The PQ pool can receive electrons from both photosynthetic and respiratory electron transport chains. In the photosynthetic electron transport chain, PSII can absorb light energy and produce electrons to be accepted by the PQ pool. In the respiratory pathway, glucose can be metabolized through glycolysis and Krebs cycle to produce NADH2, which can be dehydrogenated by the NADH dehydrogenase and donate electrons to the PQ pool. Electrons can be pulled out of PQ pool by Cyt B6f, a share point for both electron transport chains. Finally, electrons will transport to either PSI or the terminal oxidase. The redox state of PQ pool depends on the rate of electrons coming in from PSII and respiratory pathway and pulling out by PSI and terminal oxidases. The activity of PSII in H1 strain indicates that PSII can donate electrons to PQ pool at the speed at least equal to that of wild type if not faster. The conclusion can be supported by the result of DCMU rescue experiment. Since DCMU block electron transport from PSII to PQ pool, no or few electrons come from PSII under stress conditions, and PQ pool will be prevented from extremely reduced. On the other hand, rotenone can also rescue H1 strain grown under stress conditions. Since rotenone block electron transport from NADH to PQ pool, it is evident that respiratory pathway also contributes to the extreme reduction of PQ pool. We do not know exactly how inactivation of slr0110 affects redox state of PQ pool. However, since the H1 strain and psbG1 knock out mutant has the same growth phenotype, and both strains express the cryptic plasmid gene psb G2 that is silent in wild type, the function of slr0110 may be related with psb G1.

REFERENCES

**Table 1. Primers synthesized for RT-PCR**

<table>
<thead>
<tr>
<th>ORF</th>
<th>Sequence</th>
<th>Length of amplified DNA fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>slr0110A</td>
<td>AACCACAACCGCAGTTAACC</td>
<td>491</td>
</tr>
<tr>
<td>slr0110B</td>
<td>TACCATCCAACACAGACCCFC</td>
<td></td>
</tr>
<tr>
<td>slr0111A</td>
<td>AACTCCAGAGGCTTCTGTGC</td>
<td>418</td>
</tr>
<tr>
<td>slr0111B</td>
<td>CCAACCAATGCTCTTCATCGC</td>
<td></td>
</tr>
<tr>
<td>slr0112A</td>
<td>TTTACTCCTTTCAGCAACCG</td>
<td>495</td>
</tr>
<tr>
<td>slr0112B</td>
<td>CCTTCGCTGGAATAACTGG</td>
<td></td>
</tr>
<tr>
<td>slr0115A</td>
<td>AGGCTACGATGTCCAACAGG</td>
<td>555</td>
</tr>
<tr>
<td>slr0115B</td>
<td>AACTTAGGCGAGGTGACG</td>
<td></td>
</tr>
<tr>
<td>sll0094A</td>
<td>GAAGGATTAACGTTGACGCG</td>
<td>920</td>
</tr>
<tr>
<td>sll0094B</td>
<td>AGATTGGTTCATCCAGCGC</td>
<td></td>
</tr>
<tr>
<td>sll0095A</td>
<td>ATCGAACAGGAATTGCACG</td>
<td>1031</td>
</tr>
<tr>
<td>sll0095B</td>
<td>TTAACCTGTGGCAGCGACG</td>
<td></td>
</tr>
<tr>
<td>slr1841A</td>
<td>CCGACAACCTTGACGGAATGC</td>
<td>905</td>
</tr>
<tr>
<td>slr1841B</td>
<td>GGTGAACGTGACTCAAGGACC</td>
<td></td>
</tr>
<tr>
<td>sll0337A</td>
<td>TATCGTTATGGCGAACCAGG</td>
<td>503</td>
</tr>
<tr>
<td>sll0337B</td>
<td>TAATCGGTTTCTGGCTCGAGG</td>
<td></td>
</tr>
<tr>
<td>Input</td>
<td>MW (Da)</td>
<td>pl</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>h01</td>
<td>24541</td>
<td>7.08</td>
</tr>
<tr>
<td>h02</td>
<td>23861</td>
<td>5.77</td>
</tr>
<tr>
<td>h05</td>
<td>17884</td>
<td>4.78</td>
</tr>
<tr>
<td>h06</td>
<td>14231</td>
<td>4.44</td>
</tr>
<tr>
<td>w01</td>
<td>67601</td>
<td>4.47</td>
</tr>
<tr>
<td>w02</td>
<td>36009</td>
<td>4.23</td>
</tr>
<tr>
<td>w03</td>
<td>36570</td>
<td>4.57</td>
</tr>
<tr>
<td>w04</td>
<td>40023</td>
<td>4.57</td>
</tr>
<tr>
<td>w05</td>
<td>30034</td>
<td>4.45</td>
</tr>
<tr>
<td>w06</td>
<td>43910</td>
<td>4.71</td>
</tr>
<tr>
<td>w07</td>
<td>29912</td>
<td>4.82</td>
</tr>
<tr>
<td>w08</td>
<td>29912</td>
<td>4.82</td>
</tr>
<tr>
<td>w09</td>
<td>28906</td>
<td>4.95</td>
</tr>
<tr>
<td>w10</td>
<td>13239</td>
<td>5.36</td>
</tr>
<tr>
<td>w11</td>
<td>18155</td>
<td>4.98</td>
</tr>
<tr>
<td>w12</td>
<td>17355</td>
<td>4.75</td>
</tr>
<tr>
<td>w13</td>
<td>12590</td>
<td>4.96</td>
</tr>
<tr>
<td>w14</td>
<td>13239</td>
<td>5.36</td>
</tr>
<tr>
<td>w15</td>
<td>14660</td>
<td>5.2</td>
</tr>
<tr>
<td>w16</td>
<td>11135</td>
<td>5.33</td>
</tr>
</tbody>
</table>

-- No accession number in Cyanobase.
<table>
<thead>
<tr>
<th>Strain</th>
<th>ORF</th>
<th>Medium Light Deleted (40 μmol. S⁻¹.m⁻²)</th>
<th>Medium Light +Glucose</th>
<th>High Light Deleted (120 μmol. S⁻¹.m⁻²)</th>
<th>High Light +Glucose</th>
<th>Pigment content (µg. A₇₃₀⁻¹. ml⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td></td>
<td>34</td>
<td>17</td>
<td>41</td>
<td>16</td>
<td>4.4</td>
</tr>
<tr>
<td>H1</td>
<td>slr0110</td>
<td>39</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.4</td>
</tr>
<tr>
<td>H2</td>
<td>slr1841</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>H3</td>
<td>slr0408</td>
<td>33</td>
<td>16</td>
<td>40</td>
<td>15</td>
<td>4.3</td>
</tr>
<tr>
<td>H4</td>
<td>slr0151</td>
<td>33</td>
<td>15</td>
<td>39</td>
<td>15</td>
<td>4.6</td>
</tr>
<tr>
<td>H5</td>
<td>slr1358</td>
<td>32</td>
<td>15</td>
<td>38</td>
<td>14</td>
<td>4.8</td>
</tr>
<tr>
<td>H6</td>
<td>slr1270</td>
<td>33</td>
<td>15</td>
<td>39</td>
<td>16</td>
<td>5.0</td>
</tr>
<tr>
<td>H7</td>
<td>slr0168</td>
<td>32</td>
<td>16</td>
<td>38</td>
<td>15</td>
<td>4.2</td>
</tr>
<tr>
<td>H8</td>
<td>slr1668</td>
<td>37</td>
<td>16</td>
<td>40</td>
<td>14</td>
<td>4.9</td>
</tr>
<tr>
<td>H9</td>
<td>slr0431</td>
<td>32</td>
<td>15</td>
<td>37</td>
<td>16</td>
<td>4.6</td>
</tr>
<tr>
<td>H10</td>
<td>ssr2998</td>
<td>33</td>
<td>17</td>
<td>40</td>
<td>16</td>
<td>4.4</td>
</tr>
<tr>
<td>H11</td>
<td>slr0929</td>
<td>34</td>
<td>16</td>
<td>40</td>
<td>15</td>
<td>4.2</td>
</tr>
<tr>
<td>H12</td>
<td>slr1306</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

- cells died, -- no knock out mutant was created.
The figures are averages of two independent experiments.
FIGURE LEGENDS

FIG. 1. a. The construct for deleting the ORF slr0110 from *Synechocystis* sp. PCC 6803. The DNA fragment of the ORF was amplified by PCR followed by insertion into the vector pGEM-T, the cloned fragment was then digested by Hpa I to cut out a 40 bp fragment between the two HpaI sites. A spectinomycin resistance gene cassette Spec<sup>r</sup> was inserted to the vector to replace the removed 40 bp DNA fragment. b. Agrose electrophoresis of PCR products amplified from WT and H1 strain using primers specific to the ORF slr0110.

FIG. 2. The growth curve of H1 and wild type strain under different growth conditions. a. Low light intensity. b. Medium light intensity. c. High light intensity. d. Medium light intensity with the addition of DCMU in the growth medium. e. High light intensity with the addition of DCMU in the growth medium. f. Medium light intensity with the addition of rotenone in the growth medium.

FIG. 3. The cells of H1 and WT strain grown under different conditions. a. H1 and WT grown under medium light intensity without glucose. b. H1 grown under low light intensity with glucose. c. Rescue of H1 strain grown with glucose by rotenone.

FIG. 4. a. Oxygen uptake measurement for the whole WT and H1 cells, cells with 5µg chlorophyll were used for the measurement. b. Oxygen evolution measurement for the whole WT and H1 cells, 1 O.D. cells were used for the measurement.

FIG. 5. 77 K fluorescence emission spectra for cells grown under different conditions. a. cells were grown under medium light intensity without glucose. b. Cells were grown under medium light without glucose to mid-exponential phase, and then treated with 5 mM glucose for 16 hours. c. Cells were grown under medium light without glucose to mid-exponential phase for 16 hours, and then treated with 5 mM glucose and high light for 16 hours.
**FIG. 6.** Measurement of variable chlorophyll fluorescence. Cells were treated in the same way as those for 77K fluorescence emission spectra analysis.

**FIG. 7 a.** Electrophoresis of RNA from WT and H1 cells. RNA samples from two independent preparations for both WT and H1 strain were loaded in a 1.4% agarose gel. In lane WT-A, WT-B, H1-A and H1-B, RNA from same amount of cells were loaded. In all the other lanes, RNA from half amount of cells was loaded. **b.** The location of slr0110 and its neighboring ORFs in the genome. **C.** Electrophoresis of products of RT-PCR with WT and H1 RNA as templates using primers specific to the ORFs slr0110, slr0111, slr0112, slr0115, sll0094, sll0095, slr1841, sll1703 and sll0337 respectively.

**FIG. 8 a.** 2-D separation of total thylakoid membrane proteins from *Synechocystis* sp. PCC 6803 using pH range 4-7 IPG for the first dimension and 12-18% gradient SDS-PAGE for the second dimension. Thylakoid membrane proteins were isolated from **a.** WT, **b.** H1 strain. The gels were stained with colloidal Coomassie blue and analyzed with PDQuest software.

**FIG. 9** Alignment of protein sequence encoded by slr0110 and slr0929. The alignment was made by software GeneWorks.

**FIG. 10.** Proposed electron transport chains in thylakoid membrane of *Synechocystis* sp. PCC 6803.
Fig. 1

129
Fig. 2
Fig. 3
Fig. 4
**Fig. 5**

Graphs showing fluorescence in various conditions:

- **a**
  - Y-axis: Fluorescence (A.U.)
  - X-axis: Emission wavelength (nm)
  - Lines: WT NL and H1 NL

- **b**
  - Y-axis: Fluorescence (A.U.)
  - X-axis: Emission wavelength (nm)
  - Lines: WT N+G and H1 N+G

- **c**
  - Y-axis: Fluorescence (A.U.)
  - X-axis: Emission wavelength (nm)
  - Lines: WTH+G and H1 H+G
Fig. 8
Fig. 8
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>slr0110</td>
<td>GTPKAVMGQGKYTITAVLARERASVLDAPCSSPLKV 50</td>
</tr>
<tr>
<td>slr0929</td>
<td>HNHTALQGQGKYTTAVLARDQGTLLIDRANMFLA 46</td>
</tr>
<tr>
<td>Consensus</td>
<td>A...AVM...GKSGXXTTYAVL...EAPCSSPLKK 50</td>
</tr>
<tr>
<td>slr0110</td>
<td>SEALTVGALKLPFTTAPGDLILDTVKTDNKLPAHMLTLKV 141</td>
</tr>
<tr>
<td>slr0929</td>
<td>HPLIPTTAYLDPALTKEQLASLKMYKILLTKV 129</td>
</tr>
<tr>
<td>Consensus</td>
<td>E...PLTTE...E...F...TVKTDNKLPAHMLTLKV 150</td>
</tr>
<tr>
<td>slr0110</td>
<td>GLNLGETLRLIEGLIPWVFVRSYKRFEBKLDDFTQGOK 191</td>
</tr>
<tr>
<td>slr0929</td>
<td>KAFFARSEIPQLDPEFLFTTEIFLLFPARRSQSHIFDVYDPA 178</td>
</tr>
<tr>
<td>Consensus</td>
<td>E...E...E...E...E...F...F...F 200</td>
</tr>
<tr>
<td>slr0110</td>
<td>AKENDRPAVCEDDELL 211</td>
</tr>
<tr>
<td>slr0929</td>
<td>SShADDFKVRALLEL 195</td>
</tr>
</tbody>
</table>
| Consensus | ... ... ... ...... 220

Fig. 9
Mutant Phenotypes
1. Photosynthesis deficient
2. Respiration deficient
3. Light-sensitive
4. Essential for survival

Photosynthesis

Calvin Cycle

Cyt c6 or Pc

NADP+

NADPH

ADP

ATP

ATP Synthase

Respiration

NADH

NADH2

ADP

ATP

Glycolysis

Krebs Cycle

Glucose

Other Processes
1. Cofactor Synthesis and Assembly
2. Protein Targeting and Assembly
3. Regulatory Redox Light etc.
4. Other Membrane lipids, Degradation etc.

Fig. 10
CHAPTER 6. COMPUTATIONAL OPPORTUNITIES IN PROTEOMICS RESEARCH: AUTOMATION IN DATA PROCESSING AND ANALYSIS

A paper published by High Performance Computing Conference (HPC 2002), December 16-19 in Bangalore, India. Dan Harvey\textsuperscript{1}, Yingchun Wang\textsuperscript{2,3}, Parag R Chitnis\textsuperscript{2,3}, and Suresh Kothari\textsuperscript{1,3}

Application of computer programs in functional genomics research

The biological research in the 21\textsuperscript{st} century biology is expected to be a data-driven enterprise because of the rapid availability of data in all types of biological investigations. With the developments in large-scale genome sequencing, increasingly larger genomic sequence information is becoming available in the gene databases. These sequences form a launching pad for high throughput functional genomics and proteomics studies. A consequence of these activities is the production of complex multidimensional data. To obtain, organize and interpret these data, computational tools have become increasingly important in biology research.

Computation plays several roles in the high-throughput biology. First, it is required for the management of information flow in an experimental setting. Laboratory information management systems are becoming increasingly sophisticated as the data generation speed and data complexity increases. Second, specific algorithms are used to identify functional features in the sequence information. For example, computational programs are used to assemble genomic sequences, predict unknown genes, discover the genetic events during evolution process, and study the structure and function of proteins. Third, these analytical tools need to be expanded in genome-wide analysis. This is one of the most common computational needs in biology as the single gene/single protein analysis approach is being replaced by systems approaches encompassing large sets of genes and proteins.

\textsuperscript{1} Department of Electrical and Computer Engineering, Iowa State University, Ames IA, USA
\textsuperscript{2} Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames IA, USA
\textsuperscript{3} Interdepartmental Graduate Program In Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA
Computational programs can greatly reduce the time required to obtain and analyze a set of data. For each data set analyzed the user must enter the data into the program and extract the required results from the finished computation. As the size of the data set grows, the human effort can be significant for submitting one request at a time to requests the web database, waiting for the result, then parsing the result file manually to extract the relevant information, converting it to the required form, and finally aggregating the results and analyzing them. This document gives examples of how using automated batch processing can greatly reduce the associated human effort and make it possible to process thousands of requests within a short time. In this case, a computational program is used to study the proteome of a bacterial membrane system.

**Introduction to proteome and proteomics**

Proteome is the protein complement expressed by the genome of an organism [1, 2]. The goal of proteomic study is to identify putative proteins and study the expression changes, structure, interactions, and function of proteins at a genome-wide scale. When studying proteome expression, the following steps are typically used: (a) Protein isolation from cells or tissues. (b) 2-D electrophoresis to separate proteins according to isoelectric point and molecular weight. (c) Detection of protein spots and image analysis. (d) In-gel digestion with appropriate protease to generate small peptide fragments. (e) Mass spectrometry to generate peptide mass fingerprints (PMF). (f) Genomic sequence database searching to match query PMF data with a putative ORF. Computational analyses are involved in many of these steps and automation is a key component in increasing the throughput of these repetitious processes.
Once the proteins are identified, the information leads to functional genomics studies, such as the one depicted in Fig. 1. Generation of mutants and their characterization guided by the predicted features of the proteins reveals the protein function. Proteome expression information, predicted functional and structural features, and observed mutant information are integrated in a database. This typical proteomics-based functional genomics approach is being used to understand the functions of proteome of cyanobacterial thylakoids, the photosynthetic membranes that are involved in oxygenic photosynthesis [3] (Fig. 1). Two steps in the flow chart have been automated, one is the identification of protein spots from mass spectrometry data, and the other is the structure prediction for the functional study of proteome.

**Protein identification**

Several hundreds of proteins can be resolved on a single 18 cm, 2-dimensional SDS-PAGE (2-D) gel (Fig. 2). Each protein spot shown on the gel represents the product of a unique gene. To identify these proteins, the protein spots are excised and subjected to fragmentation with a protease that recognizes specific sites on the protein and cleave the protein into serial peptides with different sizes. The generated peptides are subjected to MALDI-TOF mass spectrometry, which can be performed at a high throughput manner. The output of mass spectrometric analysis reveals peptide masses (Fig. 3). Each peak on the mass-spectra represents a peptide, and its molecular weight can be measured with an error of ±25 parts per million. The peptide mass fingerprint (PMF) is the series of molecular mass data of peptides in the same spectrum. Thus, each protein spot on the gel can produce a PMF.
data file, which is used to search against a genomic database with the MS-FIT tool to identify the protein. The protein can be identified through matching the experimental PMF data with predicted PMF data, which can be obtained by theoretically cutting protein sequences with a protease (Fig. 4). The larger the number of measured peptide masses that match with the theoretical data leads to a more convincing protein identification.

### Experimental PMF
- 649.3669
- 838.4919
- 998.4429
- 1116.5740
- 1271.6740
- 1431.5502
- 1557.7853
- 1704.8311
- 2013.9517
- 2446.157
- 2775.285
- 320.3323

### Theoretical PMF
- 534.4356
- 599.7645
- 838.4920
- 998.4431
- 1001.3476
- 1271.6752
- 1704.8315
- 2013.9530
- 2446.1598
- 2775.2850
- 3320.3323

Fig. 3. The mass spectrum generated by MALDI-TOF mass spectrometry from a protein spot, each peak on the spectrum represents a peptide, the number beside each peak is the peptide mass represented by the peak.

Fig. 4. A diagram showing how to identify proteins using peptide mass fingerprinting, the experimental PMF was the data collected from the mass-spectrum, the theoretical PMF represent the mass of each peptide theoretically calculated if cutting the proteins with the same protease.

Protein identification is done using a web-based search tool called MS-Fit [4]. To identify the protein, MS-Fit needs a host of information from the user. This information includes the protein’s PMF data and associated search options. The search options include the database to search, species, range of protein’s MW, range of protein’s pi, and digest, etc. The result of the MS-Fit search returns a HTML page containing information on the identified protein. The user extracts the necessary information from the returned HTML page and adds it to a spreadsheet [5].

Thus, identifying a protein can be broken down into the following steps (assuming the PMF data for each protein has already been generated):

1. Copy a protein’s PMF data to MS-Fit.
2. Modify the search options.
3. Submit the query (Search for the protein).
4. Copy necessary information from the returned MS-Fit search result and add it to the spreadsheet.

All four steps must be repeated for each protein to be identified. It takes approximately 5 hours to identify 100 proteins on a local database using MS-Fit.

Automated batch processing for protein identification

The above process of protein identification is a repetitive and a straightforward process. Because most of the time spent identifying the protein is user overhead, doing automated batch processing could mostly eliminate this overhead. To perform automated batch processing on MS-Fit, three programs were written: msf, parse_msf, and a user front end. Figure 5 shows the interaction between the user, batch processing software, and MS-Fit.
The msf program performs the function of submitting data to MS-Fit and retrieving the results. The msf program is given a list of filenames each containing PMF data. For each PMF file given to msf, an output file is generated. The generated output file is the search result retrieved from MS-Fit for the specified PMF data. For each PMF file given to msf, the following steps are performed:

1. Submit the PMF data to MS-Fit
2. Submit the search options to MS-Fit (retrieved from the a file)
3. Instruct MS-Fit to perform the search
4. Retrieve the returned MS-Fit HTML page
5. Save the HTML page to an output file

![Screenshot of automated batch processing software front end](image)

Fig. 6. Screenshot of automated batch processing software front end

The parse_msf program parses the HTML output files generated by msf. Some of the information parsed (extracted) out of the HTML output file is rank, MOWSE score, number of masses matched, species, accession, protein name, and coverage. The program receives as input a list of HTML output files to parse and a filename to save the parsed data to. The
output file generated by parse_msf is a table in the format of a tab delineated text file (which can be imported into Excel). Each entry in the table is the results of a single protein search.

Both the msf and parse_msf programs are command line programs, and thus do not have a visual interface to them. All parameters to the programs must be passed as command line arguments. This, however, is not very convenient for the user. To solve this problem a front end was created. This front end implements the following:

1. Creates an easy to use graphical user interface for the entire process
   a. Allows the user to select a list of PMF data files for protein identification
   b. Allows the user to select a filename for receiving the results of parse_msf
   c. Generates a dialog box for modifying the search options contained in a file
   d. Provides an output window to show the current progress of the batch processing
   e. Provides buttons for starting and stopping the search.
   f. Notifies the user when finished processing

2. Runs msf and parse_msf to process the job, makes them oblivious to the user

Performance of automated batch processing for protein identification

Recall that manually submitting the data to MS-Fit and parsing the results took approximately 5 hours per 100 PMF files submitted. Using the software the same 100 PMF files would take approximately 10 minutes to submit and parse, with only a few minutes of setup time by the user. This yields a performance increase by two means. First, the user need only be at the computer for a few minutes setting up the batch job versus 5 hours doing it manually. Second, once the job has been set up it take the software roughly 10 minutes to complete the job versus 5 hours manually.

For the above searches, MS-Fit searched a local database to identify the proteins. If, however, MS-Fit were searching a remote database the time necessary to perform the search might be much higher. For the software, this increased search time only effects the time required to process the job, not to set it up. Thus, a user could set a job up in a few minutes and let the job process overnight in an unattended mode.
Prediction of trans-membrane regions of membrane proteins

Membrane proteins either span or loosely associate with lipid bilayers. About 30% of the genes in a bacterial genome are predicted to encode membrane proteins. Membrane proteins play important roles in many crucial and unique metabolic pathways. However, the study of membrane proteins has lagged behind the cytoplasmic proteins because they are embedded in lipid bilayers and thereby refractory to aqueous solutions. Therefore, the number of experimentally identified membrane proteins is far less than soluble proteins. Computational prediction is a good method to predict those unknown membrane proteins because all the membrane proteins should have one or more trans-membrane regions. Further biochemical or genetic study can be applied to study the function of those predicted membrane proteins.

Programs used to predict trans-membrane regions.

The entire genome analyzed for trans-membrane regions was *Synechocystis* sp. PCC6803. This genome consists of 3167 known or hypothetical genes. In order to carry out trans-membrane regions analysis on the genome a computational tool needs to be utilized.

There are now many programs available to predict trans-membrane regions of membrane proteins (http://www.expasy.ch). For the intended analysis above, the TMpred [6, 7] was chosen as the best program for this analysis. This is because it not only predicts the number and positions of trans-membrane helices, but also predicts the orientation of each trans-membrane helix, e.g., whether the location of N-terminal or C-terminal is inside the membrane or outside the membrane.

Manually entering each sequence into TMpred and parsing the results would be a very long and time-consuming task. In an effort to reduce this manual labor, the TMpred program was implemented with automated batch processing. To achieve this without modifying the source code of the TMpred program, a second program (a front-end) was written. The front-end requires that the user supply a list of sequence names, a database, minimum helix, maximum helix, and output filename. The steps below describe how the front-end performs the analysis for the entered information.

1. Get the first sequence name from the list
2. Search the database for the sequence name and retrieve its sequence
3. Run TMpred, supply it the minimum helix, maximum helix, and sequence
4. Extract pertinent data from the TMpred output and append it to the output file
5. Get the next sequence name from the list and jump to step 2, if one exists

The output file generated by the front-end includes the following information: protein ID and the number, locations and orientations of trans-membrane helices for each protein (table 1).

Table 1. The output format of the automated TMpred program

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Number of TM</th>
<th>ID of TM</th>
<th>Start Position</th>
<th>End Position</th>
<th>Length</th>
<th>Score</th>
<th>Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sII0008</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>30</td>
<td>20</td>
<td>2512</td>
<td>i-o</td>
</tr>
<tr>
<td>sII0010</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sII0012</td>
<td>1</td>
<td>1</td>
<td>158</td>
<td>177</td>
<td>20</td>
<td>1636</td>
<td>i-o</td>
</tr>
<tr>
<td>sII0016</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>22</td>
<td>22</td>
<td>1493</td>
<td>i-o</td>
</tr>
<tr>
<td>sII0017</td>
<td>1</td>
<td>1</td>
<td>185</td>
<td>206</td>
<td>22</td>
<td>583</td>
<td>o-i</td>
</tr>
<tr>
<td>sII0018</td>
<td>0</td>
<td></td>
<td></td>
<td>22</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sII0019</td>
<td>2</td>
<td>1</td>
<td>28</td>
<td>47</td>
<td>20</td>
<td>1089</td>
<td>o-i</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>96</td>
<td>118</td>
<td>23</td>
<td>1498</td>
<td>i-o</td>
</tr>
</tbody>
</table>

By implementing automated batch processing on the TMpred program the calculation time has been greatly reduced. The whole prediction process took about 30 minutes to run on a 300 MHz, AMD K6 machine.

From table 1, we can see that information of trans-membrane helices for a batch of proteins can be produced at the same time, the structure information of any proteins can be found easily from the output. To get the full use of the output and get the structure information of a protein in a more straightforward way, we integrated the output data into our membrane protein database and

The sequence of slr1864 is:

Red color indicates an orientation of inside to outside membrane
Blue color indicates an orientation of outside to inside membrane

```
MNSFLIIIFR LYLPSIVAIL TGULSFLSL KEQGFSNK HLSLVVPYL GKFPLWFGVP
LGQVIFVHKA NSGGLWLSL IVAAGAVVVG IFLSWLWVQT LEFFSLSLAFQ TSFLTVSTVG
NTSYGFFII LLPLQGLFDY GFWALYDLGL GFLLAAYLGL SILASRSLLLT NPFKLISFLGL
DSPVPPINQK YDTDIKSSSI KFKQMLKKN ILWTVXPNPT IAFIIGLILR PIVFPFWLDD
ILGWFPAMIS MSLILGMDM LQQLDQVHNI XIAFAVSIK MLVYPLVJAM ALTAGLSSP
FRVLLVQLSS MFSAFASLIL REAYGLDULE VTCLGCDLSL MLLTLPFWL WQPTTW
```

Fig. 7. The output generated by the program that can label trans-membrane regions of a protein directly on the sequence. The trans-membrane regions were labeled with different color according to each orientation.
use a computational program to locate the trans-membrane helices directly on the protein sequence. If we need to find the structure information of a proteins, we just need to input the protein ID to the search engine, the output will use red color to indicate an orientation of inside to outside membrane and blue to indicate outside to inside membrane, the position and length of each trans-membrane helix can be easily obtained from the labeled sequences (Fig. 7).

Using automated batch processing analysis on biological data, results can be obtained much more quickly and efficiently. This method should prove to be very useful in proteomic research.

**Conclusions**

Automation is a general trend in genomic and proteomic research, many processes remains to be automated. The secondary structure prediction of proteins, which is very similar to the trans-membrane helix prediction, can be automated and these higher order data structures could be stored in a database for analyzing homologies based on higher order structural and functional features. There are many computational programs available to predict the secondary structure for a single protein such as Prof, PSIPred, SOPM, etc, but few programs are automated to predict all protein sequences from a genome at the same time. Functional prediction of proteins can also be automated. To predict the function of an unknown protein, we can scan the protein sequence for the known functional motifs and domains, and then link the function of the unknown protein to the known proteins with the same motifs and domains. Many databases for functional domain and motif exist such as PROSITE, Pfam and PRINTS, and also many programs exist to search a protein sequence against the databases to predict the functional motifs and domains, but the automated program which can scan a whole proteome sequence against databases is still unavailable.

Automation in computational processes is essential for increasing the throughput of data gathering and analysis. This is demonstrated well at the DNA sequence level in the large-scale genome projects. However, functional genomics research is currently at the technology development phase and as the technologies of proteomics, metabolomics and phenomics become standardized, automation will be possible for efficient data collection and interpretation.
Reference


3. Wang Y, Sun J, Chitnis PR. Proteomic study of the peripheral proteins from thylakoid membranes of the cyanobacterium Synechocystis sp. PCC 6803 Electrophoresis 2000 May;21(9):1746-54


CHAPTER 7. THE MEMBRANE PROTEOME DATABASE

1. Introduction

Two-dimensional electrophoresis (2-DE) in conjunction with computer analysis is one of the best approaches to perform global analysis of protein expression in complex biological samples. Web-based 2-DE protein databases have been created and proven to be useful for data exploit and interchange via internet [1, 2]. However, no database has been created specifically for membrane proteins. In our group, thylakoid membrane proteins from Synechocystis sp. PCC 6803 have been resolved in 2-D gels and identified by peptide mass fingerprinting. We tried to construct a web accessible database that includes the reference 2-D maps as well as the structure and functional information of membrane proteins.

2. Database design

Five entities will be included in the database, as shown in the ER-diagram (Figure 1) Five entities will be included in the database, as shown in the ER-diagram (Figure 1)
The entity **Proteins** includes all identified proteins resolved on 2-D gels, each protein has the attributes P_ID (accession number in CyanoBase), P_Name (protein name), GRAVY value, TM (number of transmembrane helices), Mass, pI and NCBI_ID (accession number in NCBI database). The entity **2D_Gels** includes all identified 2-D gels; each 2-D gel will have the attributes Gel_ID and Description. The Description attribute includes related information of the 2-D gel such as pH range of first dimension and the concentration gradient of the second dimension as well as the protein sample information. The entity **Protein_Spots** includes all the identified protein spots, each protein spot must be on a unique 2-D gel, and has the attributes Spot_ID, Protein_ID and Gel_ID. Each protein spot identified on 2-D gels must be encoded by a unique gene from Synechocystis genome. The entity **Domain_motif** includes all the possible functional domains and motifs predicted from the Synechocystis proteome; each domain and motif must belong to one or more proteins. The entity **Mutants** includes the information of all of the generated mutants of the Synechocystis; each mutant must have one gene mutated, and the phenotype attribute will include a brief description of the phenotype of the mutant. The person who generated the mutant will also be recorded.
The schema of the database is:

**Proteins** (P_ID, P_Name, GRAVY, TM, Mass, pI, NCBI_ID),

**2D_Gels** (Gel_ID, Description),

**Proteins_Spots** (Spot_ID, Gel_ID, Protein_ID, Protein_ID is the foreign key to Proteins, Gel_ID is a foreign key to 2D_Gels),

**Mutants** (P_ID, Gene, Investigator, Phenotype, P_ID is the foreign key to Proteins),

**Domain_motif** (D_ID, D_Name, P_ID, P_ID is the foreign key to Proteins).

No partial dependence exists nor non-key attribute is transitively dependent on the key. Therefore, the relation schema is in 3-NF and redundancy is avoided. Since many protein sequences contain more than 256 amino acid residues, which are beyond the allowed
size of an attribute of the MySQL database, we just save the protein sequence translated by
the whole genome as a plain text file.

3. Data collection

**Hydrophobicity prediction:** To evaluate the hydrophobicity of a protein, we use two
routine criteria. One is to predict the overall hydropathy of the protein, which is known as the
GRAVY value, and the other is to predict the number of transmembrane helices a protein has.
To predict the GRAVY value of a protein, we assign a value to each amino acid according to
the hydrophobicity of its side chain, and then we calculate the average value of all of the
amino acid residues in the protein [3]. GRAVY value may not reflect the true hydropathy of
a protein because some membrane proteins contain short transmembrane helices but have a
long cytoplasmic loop, which may have a low GRAVY value. But from the genome scale
level, we found the trend that the number of transmembrane domains increased with an
increase in the GRAVY value. We developed a program with C++ to calculate the GRAVY
value for all the proteins encoded by the whole genome of *Synechocystis*. To predict the
number of transmembrane helices a protein has, we use the program **TMpred** and automated
it to predict the proteins encoded by the whole genome in a throughput way [4]. This
program not only predicts the number and positions of transmembrane helices, but also
predicts their orientations.

**Functional motif and domain prediction:** We can use the program InterPro_Scan to
predict the functional domain and motifs from a protein sequence [5]. This program scans a
protein sequence against the PROSITE, Pfam, PRINTS and other family and domain
databases.

4. Database construction and application development

A web-based database for membrane proteins was constructed on a WINDOWS
based Apache server. MySQL was used to create the database; HTML, SQL, PERL and
JAVASCRIPT were used to develop user applications. Reference maps of thylakoid
proteome of *Synechocystis* as well as part of structure information of membrane proteins
such as GRAVY value and number of transmembrane helices have been included. In the
home page of the database, all the user and manager applications are listed in the left frame;
we can click on each links to perform corresponding actions to the database. All the applications will be described in the following paragraphs. In the mainframe of the homepage, users can input their suggestions and comments in the form, and send them to the webmaster via email.

**The reference 2-D protein maps**  The database currently includes 6 2-D reference maps for thylakoid membrane proteins of *Synechocystis*. The gels are different by the first dimension pH range (for example, pH 3-10, pH 4-7 and etc.) and the proteins resolved (peripheral and integral proteins), as shown by Fig. 2. All the analyzed spots in each gel were labeled with a unique number, some of the labeled spots were identified by PMF analysis and the others were not due to the reasons as we discussed [6]. The ID of each identified protein spots can be found by searching the database using the labeling number. Fig. 2 shows that there is a search engine in the bottom frame. To find the ID and the related information of a protein such as the number 4 in the peripheral fraction resolved in 2-D gel with pH range 3-10, we first need to select the corresponding gel in the dropdown box *Peripheral pH 3-10*, and then input number 4 in the query box, press the submit button to search against the database. The protein ID and its structural information will be displayed in the bottom frame (Figure 3). The information includes the protein name, Cyanobase number,
NCBI accession number, deduced mass, deduced pl, gravy value and the number of transmembrane helices. The sequence of the protein is also given and labeled with the location and orientation of transmembrane helices. Note that the Cyanobase and NCBI number each has a link on it. If we click the link, then the related information of the protein will be retrieved from the two respective databases. The cross links to foreign databases make the database more dynamic and the users can easily get the up-to-date information of the protein. The information of the protein spots in other gels can also be found in the same way.

**Search the database** If a user wants to know whether a specific protein was identified or not in the database, the only thing need to do is to input one of the following:

- Cyanobase accession number,
- NCBI number or protein name to query the database. As shown in Figure 4, we first need to select the type of the keyword in the dropdown box, and then input the keyword in the query box. For example, if we select the type of keyword as **Protein Name** and enter the key word ATP, then all the identified proteins whose name contains the string ATP will be listed in a table, as shown in the Fig. 5. The protein name, Cyanobase number and NCBI accession number all have a link on it. The structure information of the protein can be obtained from the local database if we click the link on the protein name. Also we can click the link on the NCBI accession number or Cyanobase accession number to retrieve more related information from...
the two databases. The database also provides a search engine to search the structure information of any proteins of *Synechocystis*, as shown by the search box in the low part of Fig. 4, but can just search with Cyanobase accession number. For example, if we want to know the topology information and the gravy value of the protein encoded by the ORF slr1841, we can input the accession number slr1841 and search against the database.

**Compare the identified proteins** Different research groups may conduct proteome analysis for the same organism, and different proteins may be identified by different approaches in different groups. For example, we totally identified 128 proteins, 87 of them were not identified in Cyano2Dbase. If we try to find out which proteins were identified by both groups by comparing the identified proteins from both groups one by one, it will be boring and time costing. The database provides a function to compare the identification work in batch. Fig. 6 shows that there are two query boxes provided by the database. Users can either load the file that containing the Cyanobase accession numbers to the upper search box or just input all of the accession numbers in the bottom box, and then press the button *Compare* to find all proteins identified by both.

**List identified proteins by gel** Identified proteins from a specific 2-D gel can be displayed in a table, providing an easy way for users to download and analyze the identified proteins in a 2-D gel. In the homepage of the database, we can click the link *List identified proteins by gel*, a dropdown box will appear in the main frame, we can select any of the gels from the dropdown box and press the submit button, all the identified proteins will be listed in a table, as shown in Figure 7. The entries in the table can be sorted by each column. For example, if we click the column name Spot_ID, the tables will be sorted according to the number labeled on the gel. Similarly, we can sort the entries in the table by protein mass, pI, gravy value, number of transmembrane helices and etc. Again, the links on
the Cyanobase number and the NCBI accession number can be clicked to cross link to respective foreign database.

<table>
<thead>
<tr>
<th>The proteins identified in the 2-D gel P.4.5:</th>
<th>inputID</th>
<th>Spot ID</th>
<th>name</th>
<th>cyano ID</th>
<th>NCBI ID</th>
<th>mass</th>
<th>pI</th>
<th>gravy</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>1</td>
<td>Chloroplast import-associated channel IAP75</td>
<td>8kr1227</td>
<td>1652591</td>
<td>92307.8 4.29 -0.225</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p03</td>
<td>2</td>
<td>General secretion pathway protein D</td>
<td>8kr1277</td>
<td>1653364</td>
<td>84642.3 4.64 -0.107</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p04</td>
<td>3</td>
<td>probable porin, major outer membrane protein</td>
<td>8kr1841</td>
<td>1652528</td>
<td>67601.3 4.47 -0.147</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p04</td>
<td>4</td>
<td>probable porin, major outer membrane protein</td>
<td>8kr1841</td>
<td>1652528</td>
<td>67601.3 4.47 -0.147</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p04</td>
<td>5</td>
<td>probable porin, major outer membrane protein</td>
<td>8kr1841</td>
<td>1652528</td>
<td>67601.3 4.47 -0.147</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p12</td>
<td>6</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>a120408</td>
<td>1001111</td>
<td>43909.7 4.71 -0.039</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p12</td>
<td>7</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>a120408</td>
<td>1001111</td>
<td>43909.7 4.71 -0.039</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p12</td>
<td>8</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>a120408</td>
<td>1001111</td>
<td>43909.7 4.71 -0.039</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p12</td>
<td>9</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>a120408</td>
<td>1001111</td>
<td>43909.7 4.71 -0.039</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p24</td>
<td>10</td>
<td>putative homolog of plant HCF136, which is essential for stability or assembly of photosystem II (ycf48)</td>
<td>8kr2034</td>
<td>1652167</td>
<td>37291.1 4.76 -0.234</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p20</td>
<td>11</td>
<td>Iron transport protein</td>
<td>8kr1295</td>
<td>1651916</td>
<td>39370.4 4.87 -0.295</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p21</td>
<td>12</td>
<td>Phosphofructokinase</td>
<td>8kr11196</td>
<td>3212297</td>
<td>358588.4 5.87 0.043</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p39</td>
<td>13</td>
<td>Photosystem II manganese-stabilizing polypeptide</td>
<td>a110427</td>
<td>131387</td>
<td>29911.8 4.82 -0.207</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p25</td>
<td>14</td>
<td>30S ribosomal protein S1</td>
<td>8kr12356</td>
<td>2500385</td>
<td>35670.2 4.57 -0.306</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p39</td>
<td>15</td>
<td>Photosystem II manganese-stabilizing polypeptide</td>
<td>a110427</td>
<td>131387</td>
<td>29911.8 4.82 -0.207</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p39</td>
<td>16</td>
<td>Photosystem II manganese-stabilizing polypeptide</td>
<td>a110427</td>
<td>131387</td>
<td>29911.8 4.82 -0.207</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. List of all the identified proteins in the 2-D gel for the peripheral proteins with pH range 4-5 for the first dimension

Maintain the database

The manager of the database can add, delete or update information in the database. For security purpose, the manager must input his username and password before performing any actions to the database. If the username and the password are correct, then a table containing the name of tables in the database and the actions will be displayed, as shown in Figure 8. In the left column of the table, we can check the radio button of one table name that need to be
modified, and select the corresponding action in the right column. For example, if we want to add a new identified protein to the table *total identified proteins*, then we can select the table and the *ADD* action. After submit the selection of table and action, a new table will be displayed, as shown in Figure 9, we can add related information to the table. For example, the *InputID*, *Protein Name*, *Deduced mass*, *Deduced pI*, *Cyanobase No.* and *NCBI No.*. Similarly, we can perform other actions to all the other tables.

**View the contents of the whole database** The database provides a method for users to download all its contents. If we click on the link on *View the contents of the database* in the left frame of the homepage, a table that contains the names of all tables in the database will be displayed (Fig. 10). All the other tables we described are presented here as well as several other tables that we kept just for the purpose of data management. If we click the link on any of the table name, the contents of the table will be displayed. For example, if we click the table 2d_gels, a table containing all 2-D gels and their description will be displayed. Users can download all of the data in the table by simply copy and paste. The data in each table can be sorted by each column. For example, in the table *total_identified_proteins*, we can sort data according to the protein mass, pI, accession number and etc. The data in all the other tables can be sorted and downloaded in the same way.

5. Conclusion

The web based database for thylakoid membrane proteome of *Synechocystis* sp. PCC 6803 has been created. The database has user-friend interfaces and is easy for routine
maintenance. The database can cross link to foreign database easily to retrieve more information related to the proteins it contains. The database is very useful for those proteome researchers working on the same organism by providing different searching and comparison engine. With more structure and functional information available in the future, the information in the database could be used to predict the homology between proteins based on secondary structural information.

References

CHAPTER 8. GENERAL CONCLUSIONS

The proteome of thylakoid membrane of *Synechocystis* sp. PCC 6803 is estimated to contain 600-1000 proteins. Many of the proteins play important roles in photosynthesis, respiration and other metabolic pathways. 2-DE in conjunction with mass spectrometry provides a powerful method to resolve and identify hundreds of proteins at the same time. However, as membrane proteins are refractory to aqueous solutions, it is difficult to solubilize membrane proteins for successful separation with 2-D gels. When we extracted peripheral proteins from the thylakoid membrane and resolve them on a 2-D gel with pH range 3-10, more than 200 protein spots were detected, and 78 spots were identified as protein products of 52 individual genes. In order to resolve more membrane proteins, especially integral proteins, we employed serial extraction to separate integral proteins from peripheral proteins, and proteins from the two fractions were resolved on 2-D gels with different pH range. Additional peripheral and integral proteins were resolved and identified. Totally, we identified 128 proteins including 38 hypothetical proteins that have unknown functions.

To study the function of identified novel proteins, we adopted mutagenetic analysis. Knockout mutants of 12 open reading frames (ORFs) that encode hypothetical proteins were created. One of these mutants (H1), which has a deletion in the ORF slr0110, is sensitive to glucose and light. Physiological analysis of this mutant indicated that the PQ pool in the thylakoid membranes of the mutant may be extremely reduced, causing its sensitivity to light and glucose. RT-PCR analysis indicated that the expression of the ORFs downstream to the ORF slr0110 was highly increased. 2-DE analysis showed that a plasmid gene psbG2, which is only expressed in the photomixotrophic sensitive psbG1 deletion mutant, is also expressed in the H1 mutant. Therefore, we believe that function of the hypothetical protein encoded by the ORF slr0110 is related with psbG1 gene.

To best use the proteome data and to facilitate inter-laboratory data exchange and comparison, we constructed a web based relational database that includes 2-D reference maps for thylakoid membrane proteins and proteins structural and functional information. Users can compare their 2-D maps with the 2-D maps in the database, and find identified
proteins and related information via internet. The database can be maintained by a manager without knowing its internal structure.

The work included in this thesis gave an example on how to study the functions of a subproteome. With available genomic sequence of more organisms in the near future, the strategy used in this thesis could be applied to the global functional analysis of the proteome of these organisms. The database we constructed will be very useful for those who conduct research using the same organism as the model system, and the structure of the database can also be used to construct databases for the other specific proteomes.
APPENDIX A. PROTEINS OF THE CYANOBACTERIAL PHOTOSYSTEM I

Abstract from a paper published in the Biochimica et Biophysica Acta

Wu Xu, Huadong Tang, Yingchun Wang and Parag R. Chitnis

Abstract

Cyanobacterial photosystem I is remarkably similar to its counterpart in the chloroplast of plants and algae. Therefore, it has served as a prototype for the type I reaction centers of photosynthesis. Cyanobacterial PS I contains 11-12 proteins. Some of the cyanobacterial proteins are modified post-translationally. Reverse genetics has been used to generate subunits-deficient cyanobacterial mutants, phenotypes of which have revealed the functions of the missing proteins. The cyanobacterial photosystem I proteins bind cofactors, provide docking sites for electron transfer proteins, participate in tertiary and quaternary organization of the complex and protect the electron transfer centers. Many of these mutants are being used in sophisticated structure-function analyses. Yet, the roles of some proteins of the cyanobacterial photosystem I are unknown. It is necessary to examine functions of these proteins on a global scale of cell physiology, biogenesis and evolution.
APPENDIX B. THERMODYNAMICS OF ELECTRON TRANSFER IN OXYGENIC PHOTOSYNTHETIC REACTION CENTERS: A PULSED PHOTOACOUSTIC STUDY OF ELECTRON TRANSFER IN PHOTOSYSTEM I REVEALS A SIMILARITY TO BACTERIAL REACTION CENTERS IN BOTH VOLUME CHANGE AND ENTROPY

Abstract from a paper published in Biochemistry
Jian-Min Hou, Vladimir A. Boichenko, Ying-Chun Wang, Parag R. Chitnis, and David Mauzerall

Abstract
The thermodynamic properties of electron transfer in biological systems are far less known in comparison with that of their kinetics. In this paper the enthalpy and entropy of electron transfer in the purified photosystem I trimer complexes from Synechocystis sp. PCC 6803 have been studied, using pulsed time-resolved photoacoustics on the 1 micros time scale. The volume contraction of reaction centers of photosystem I, which results directly from the light-induced charge separation forming $P_{700}^{-}\tilde{F}(A)/\tilde{F}(B')$ from the excited-state $P_{700}^*$, is determined to be $-26 +/- 2 \text{ A}^3$. The enthalpy of the above electron-transfer reaction is found to be $-0.39 +/- 0.1 \text{ eV}$. Photoacoustic estimation of the quantum yield of photochemistry in the purified photosystem I trimer complex showed it to be close to unity. Taking the free energy of the above reaction as the difference of their redox potentials in situ allows us to calculate an apparent entropy change ($T\Delta S$) of $+0.35 +/- 0.1 \text{ eV}$. These values of $\Delta V$ and $T\Delta S$ are similar to those of bacterial reaction centers. The unexpected sign of entropy of electron transfer is tentatively assigned, as in the bacterial case, to the escape of counterions from the surface of the particles. The apparent entropy change of electron transfer in biological system is significant and cannot be neglected.
APPENDIX C. THE SOURCE CODE FOR APPLICATION PROGRAMS IN THE DATABASE

C1. The source code for the program database_browse.pl

```perl
#!/c:/perl/bin/perl
#
# This is program browse the contents of the database. Each table in the database will be
# shown, and the contents in the table can be displayed and sorted. The Cyanobase accession
# number and NCBI accession number of each protein can be clicked to retrieve related
# information from the two databases.
#
use CGI;
use DBI;
use strict;

my ($cgi)= new CGI;
my ($host_name)= undef;
my ($user_name)= undef;
my ($password)= undef;
my ($db_name)= "cyano_membrane_protein";
my ($dbh, $sth);
my ($dsn)= "DBI:mysql:$db_name";
$dsn .= ":hostname=$host_name" if $host_name;
$dsn .= ";mysql_read_default_file=/usr/local/apache/conf/cyano_membrane_protein.cnf";
my (%attr) = (RaiseError=> 1 ) ;
$dbh=DBI->connect($dsn, $user_name, $password, %attr);

my ($title) = "$db_name Database Browser";

# Print the output in HTML
print $cgi->header();
print $cgi->start_html(-title=>$title);
print $cgi->hl($title);

my($tbl_name)=$cgi->param("tbl_name");
my($sort_column)=$cgi->param("sort_column");
if(!$tbl_name){
    display_table_list()#Display all the tables in the database.
}
else{
    display_table($tbl_name, $sort_column);#Display a contents of a table.
}
print $cgi->end_html();

#The subroutine for displaying all the tables in the database
sub display_table_list
{
    my($ary_ref, $url);
}
Select a table by clicking on its name: 

```
$ary_ref=$dbh->selectcol_arrayref(qq{SHOW TABLES FROM $db_name});
print "<TABLE BORDER>\n";
print "<TR>\n";
display_cell("TH", "Table Name", 1);
print "<TR>\n";
foreach my $tbl_name (@{$ary_ref})
{
    $url=$cgi->script_name();
    $url.=sprintf("?tbl_name=%s", $cgi->escape($tbl_name));
    print "<TR>";
    display_cell("TD", $cgi->a({-href=>$url}, $tbl_name), 0);
    print "</TR>\n";
}
print "</TABLE>\n";
```

# The subroutine for displaying a table
sub display_table
{
    my ($tbl_name, $sort_column) = @_; 
    my ($sth, $url);
    # if sort column not specified, use first column
    $sort_column = "1" unless $sort_column;
    print $cgi->a ({-href => $cgi->script_name()}, "Show Table List");
    print "<BR><BR>\n";
    $sth = $dbh->prepare (qq{SELECT * FROM $tbl_name ORDER BY $sort_column});
    $sth->execute ();
    print "<B>Contents of $tbl_name table: </B><BR><BR>\n";
    # display table with a border
    print "<TABLE BORDER>\n";
    # use column names for table headings; make each heading a link
    # that sorts output on the corresponding column
    print "<TR>\n";
    my (@ary2);
    foreach my $col_name (@{$sth->{NAME}})
    {
        push(@ary2, $col_name);
        $url = $cgi->script_name();
        $url.= sprintf("?db_name=%s", $cgi->escape($db_name));
        $url.= sprintf("?tbl_name=%s", $cgi->escape($tbl_name));
        $url.= sprintf("&sort_column=%s", $cgi->escape($col_name));
        display_cell("TH", $cgi->a({-href => $url}, $col_name), 0);
    }
    print "</TR>\n";
    print "</TABLE>\n";
    my(@ary2)=@{$sth->{NAME}};
    # display table rows
    while (my @ary = $sth->fetchrow_array())
    {
        print "<TR>\n";
    }
    #my(@ary2)=@{$sth->{NAME}};
}
```
The following code displays each row, for cyanobase accession number and NCBI accession number, links will be made to retrieve related information from the two databases.

```perl
for ( my $i=0; $i<scalar(@ary);$i++)
{
    if($ary2[$i] eq "ID_in_NCBI" || $ary2[$i] eq "NCBI_ID")
    {
        $url2.="&db=protein&list_uids=$ary[$i]&dopt=GenPept";
        display_cell("TD", $cgi->a({-href => $url2}, $ary[$i]), 0);
    }
    elsif ($ary2[$i] eq "proteinID" || $ary2[$i] eq "P_ID" || $ary2[$i] eq "cyano_ID")
    {
        my($url3)="http://www.kazusa.or.jp/cyanobase/Synechocystis/cgi-bin/";
        $url3 .="geinfo.cgi?type=orf&name=$ary[$i]";
        display_cell("TD", $cgi->a({-href => $url3}, $ary[$i]), 0);
    }
    else
    {
        display_cell("TD", $ary[$i], 1);
    }
}
print "</TR>\n";
}

$sth->finish();
print "</TABLE>\n";
```

# A subroutine for displaying each cell
sub display_cell
{
    my ($tag, $value, $encode) = @_; 
    $value = $cgi->escapeHTML ($value) if $encode; 
    $value = "&nbsp;" unless $value; 
    print "<$tag>$value</$tag>\n"; 
}

C2. The source code for the program searches.pl

```perl
#!/c:\perl\bin\perl
# This is program browse the contents of the database. Each table in the database will be
# shown, and the contents in the table can be displayed and sorted. The Cyanobase accession
# number and NCBI accession number of each protein can be clicked to retrieve related
# information from the two databases.
```
use CGI;
use DBI;
use strict;

my ($cgi)= new CGI;
my ($host_name)= undef;
my ($user_name)= undef;
my ($password)= undef;
my ($db_name)= "cyano_membrane_protein";
my ($dbh, $sth);
my ($dsn)= "DBI:mysql:$db_name";
$dsn.= ":hostname=$host_name"if $host_name;
$dsn .= ";mysql_read_default_file=/usr/local/apache/conf/cyano_membrane_protein.cnf";
my(%attr)=(RaiseError=>1);
$dbh=DBI->connect($dsn, $user_name, $password, %attr);
my ($title)= "$db_name Database Browser";

# Print the output in HTML
print $cgi->header();
print $cgi->start_html(-title=>$title);
print $cgi->hl (Stitle);
my($tbl_name)=$cgi->param("tbl_name");
my($sort_column)=$cgi->param("sort_column");
if(!$tbl_name)
{
    display_table_list();#Display all the tables in the database.
}
else
{
    display_table($tbl_name, $sort_column);#Display a contents of a table.
}
print $cgi->end_html();
#*****************************************************************************
#The subroutine for displaying all the tables in the database
sub display_table_list
{
    my($ary_ref, $url);
    print "Select a table by clicking on its name: \n";
    $ary_ref=$dbh->selectcol_arrayref(qq{SHOW TABLES FROM $db_name});
    print "\n";
    print "\n";
    display_cell("TH", "Table Name", 1);
    print "\n";
    foreach my $tbl_name (@{$ary_ref})
    {
        $url=$cgi->script_name() ;
        $url.= sprintf("?tbl_name=%s", $cgi->escape($tbl_name));
        print "<TR><A HREF="$url">$tbl_name</A></TR>\n";
    }
}
#*****************************************************************************
print "</TABLE>\n";
}

# The subroutine for displaying a table
sub display_table
{
  my ($tbl_name, $sort_column) = @_;  
  my ($sth, $url);
  # if sort column not specified, use first column
  $sort_column = "1" unless $sort_column;
  print $cgi->a ({-href => $cgi->script_name()}, "Show Table List");
  print "<BR><BR>
";
  $sth = $dbh->prepare ("SELECT * FROM $tbl_name ORDER BY $sort_column");
  $sth->execute ();
  print "<B>Contents of $tbl_name table:</B><BR>
  # display table with a border
  print "<TABLE BORDER>\n";
  # use column names for table headings; make each heading a link
  # that sorts output on the corresponding column
  print "<TR>
";
  my (@ary2);
  foreach my $col_name (@{$sth->{NAME}})
  {
    push(@ary2, $col_name);
    $url = $cgi->script_name ();
    #$url =~ sprintf("%db_name=%s", $cgi->escape($db_name));
    $url =~ sprintf("%tbl_name=%s", $cgi->escape($tbl_name));
    $url =~ sprintf("%sort_column=%s", $cgi->escape($col_name));
    display_cell ("TH", $cgi->a ({-href => $url}, $col_name), 0);
  }
  print "</TR>\n";
  my(@ary2)=@{$sth->{NAME}};
  # display table rows
  while (my @ary = $sth->fetchrow_array ())
  {
    print "<TR>
";
    # The following code display each rows, for cyanobase accession number and NCBI accession number, # links will be made to retrieve related information from the two databases.
    for ( my $i=0; $i< scalar(@ary); $i++)
    {
      if ($ary2[$i] eq "ID_in_NCBI" || $ary2[$i] eq "NCBI_ID")
      {
        $url2.="&db=protein&list_uids=$ary[$i]&dopt=GenPept";
        display_cell("TD", $cgi->a ({-href => $url2}, $ary[$i]), 0);
      }
      elsif ($ary2[$i] eq "proteinID" || $ary2[$i] eq "P_ID" || $ary2[$i] eq "cyano_ID")
      {
        my($url3)="http://www.kazusa.or.jp/cyanobase/Synechocystis/cgi-bin";
      }
A subroutine for displaying each cell
sub display_cell
{
    my ($tag, $value, $encode) = @_;  
    $value = $cgi->escapeHTML ($value) if $encode;  
    $value = "" unless $value;  
    print "<$tag>$value</$tag>";
}

# C3. The source code for the program list_by_gel.pl

#!/c:/perl/bin/perl
#--------------------------------------------------------------
# This program provides a function to display all the identified proteins in a specific 2-D
# gel as a table, also the related information of the proteins will be displayed. Users only
# need to select the 2-D gel from a dropdown box in the web browser, and submit the selection
# to display all the proteins.
#--------------------------------------------------------------
use CGI;
use DBI;
use strict;
use FileHandle;

my ($cgi)= new CGI;
my ($dsn)= "DBI:mysql:cyano_membrane_protein:localhost";
my ($user_name)= undef;
my ($password)= undef;
my ($dbh, $sth1);
my (@ary);
my (@ary2); # Hold the name of columns
my ($sth1);

my ($gel_ID)= $cgi->param("gel_ID");
my($sort_column)=$cgi->param("sort_column");
print $cgi->header();
print $cgi->start_html ( -title => "Searching ID of a protein on 2D gel");
print "<pre>
print $cgi->h3 ("The proteins identified in the 2-D gel $gel_ID:");

#connect to database
$dbh=DBI->connect($dsn, $user_name, $password, {RaiseError=>1});

#To find out how many proteins were identified in the 2-D gel.
$sth1=$dbh->prepare(qq { select count(*) from total_identified_proteins p, tmhelix t, protein_spots s where (s.Gel_ID-"$gel_ID") and (s.cyano_ID=p.cyano_ID) and (p.cyano_ID=t.proteinID) });

$sth1->execute();
my (@ary3) = $sth1->fetchrow_array();
my ($count)=@ary3[0];
print_table($sort_column, $gel_ID); #print the searched results in a table
print "<br>
print $cgi->h3("Totally <font color='red'>$count</font> protein spots were identified in the 2-D gel $gel_ID.
<br>

$sth1->finish();
$dbh->disconnect();

#The subroutine is used to print a table
sub print_table
{
    my ($sort_column, $gel_ID)=@_;
    $sort_column = "1" unless $sort_column;
    print "<TABLE BORDER>\n";
    #Print the name of each column of a table
    print "<TR>\n"
    #Select all the identified proteins in the 2-D gel and all of the related information of the proteins.
    my ($sth)=$dbh->prepare(qq { select p.inputID, s.Spot_ID, p.name, p.cyano_ID, p.NCBI_ID, p.mass, p.pl, t.gravay, t.TM from total_identified_proteins p, tmhelix t, protein_spots s where (s.Gel_ID-"$gel_ID") and (s.cyano_ID=p.cyano_ID) and (p.cyano_ID=t.proteinID) ORDER BY $sort_column });
    $sth->execute();
    foreach my $col_name (@{$sth->{NAME}})
    {
        push(@ary2, $col_name);
        # Add a link to each column name.
my ($url) = $cgi->script_name ();
$url = sprintf ('?gel_ID=$gel_ID');
$url = sprintf ('&sort_column=%s', $cgi->escape ($col_name));
display_cell ('TH', $cgi->a ({-href => $url}, $col_name), 0);
#
#display_cell ('TH', $col_name, 0);
}
print '</TR>
';
#
# display table rows
while (my @ary = $sth->fetchrow_array ())
{
  print '<TR>
';
  for ( my $i=0;$i<scalar(@ary);$i++)
  { 
    if (Sary2[$i] eq "ID_in_NCBI" || Sary2[$i] eq "NCBI_ID")
    {
      $url2.="&db=protein&list_uids=$ary[$i]&dopt=GenPept";
      display_cell("TD", $cgi->a({-href => $url2}, $ary[$i]), 0);
    }
    elsif ($ary2[$i] eq "proteinID" || $ary2[$i] eq "PID" || $ary2[$i] eq "cyano_ID")
    {
      my($url3)="http://www.kazusa.or.jp/cyanobase/Synechocystis/cgi-bin/";
      $url3 .= "geinfo.cgi?type=orf&name=$ary[$i];
      display_cell("TD", $cgi->a({-href => $url3}, $ary[$i]), 0);
    }
    else
    { 
      display_cell("TD", $ary[$i], 1);
    }
  }
  print '</TR>
';
}
print '</TABLE BORDER>
';
$sth->finish();
}
sub display_cell
{
  my ($tag, $value, $encode) = @_;
  $value = $cgi->escapeHTML ($value) if $encode;
  $value = "0" unless $value;
  print "<$tag>$value</$tag>
";
}

C4. The source code for the program login.pl
#! c:\perl\bin\perl
#This program is developed for the maintenance of the database. the users need to login before
#performing any actions to the database. All the interfaces for ADD, DELETE and UPDATE actions
use CGI;
use strict;
use FileHandle;

my ($cgi)=new CGI;
my ($line);
my (@ary);
my (@user);
my (%users);
my ($url)=$cgi->script_name();
my ($url1)=$cgi->script_name();
my ($username)=$cgi->param("username");
my ($password)=$cgi->param("password");

print $cgi->header();
print $cgi->start_html();
print "<pre>";
chdir("..");
my ($fd)=new FileHandle("user_pass.txt")||die "Failed to open the file for usernames\n";
while($line=<$fd>)
{
  chomp($line);
  @ary=split('t', $line);
  push(@user, @ary);
  @ary=();
}
%users=@user;

#If the username and password are correct, display the table names in the database and the allowed actions.
if ($cgi->param("table") eq "" and $cgi->param("action") eq "")
{
  if (exists($users{$username}))
  {
    if ($users{$username} eq $password)
    {
      print "<font size=4 face=Verdana color=0099FF> Please select tables and
actions:<font size=4 face=Verdana color=0099FF>\n";
      print "<table border=1 width=50% bgcolor=beige>";
      print <<<FORM;
      <form action='login.pl' method=post>
      <th align=left>Tables</th><th align=left>Action</th>
      <tr><td><input type=radio name=table value=2d_gels>2-D gels</td><td rowspan=2><input type=radio name=action value=add>ADD</td></tr>
      <tr><td><input type=radio name=table value=domain_motif>domain and
motif</td></tr>
      <tr><td><input type=radio name=table value=total_identified_proteins>total
identified proteins</td><td rowspan=2><input type=radio name=action value=delete>DELETE</td></tr>
      <tr><td><input type=radio name=table value=mutants>mutants</td><td rowspan=2><input type=radio name=action value=update>UPDATE</td></tr>
      <tr><td><input type=radio name=table value=protein_spots>protein
spots</td><td><input type=submit value=Submit'></td></tr>
      </form>
    
  
  
# are provided by the program.
# -----------------------------------------------------------------------------
use CGI;
use strict;
use FileHandle;

my ($cgi)=new CGI;
my ($line);
my (@ary);
my (@user);
my (%users);
my ($url)=$cgi->script_name();
my ($url1)=$cgi->script_name();
my ($username)=$cgi->param("username");
my ($password)=$cgi->param("password");

print $cgi->header();
print $cgi->start_html();
print "<pre>";
chdir("..");
my ($fd)=new FileHandle("user_pass.txt")||die "Failed to open the file for usernames\n";
while($line=<$fd>)
{
  chomp($line);
  @ary=split('t', $line);
  push(@user, @ary);
  @ary=();
}
%users=@user;

#If the username and password are correct, display the table names in the database and the allowed actions.
if ($cgi->param("table") eq "" and $cgi->param("action") eq "")
{
  if (exists($users{$username}))
  {
    if ($users{$username} eq $password)
    {
      print "<font size=4 face=Verdana color=0099FF> Please select tables and
actions:<font size=4 face=Verdana color=0099FF>\n";
      print "<table border=1 width=50% bgcolor=beige>";
      print <<<FORM;
      <form action='login.pl' method=post>
      <th align=left>Tables</th><th align=left>Action</th>
      <tr><td><input type=radio name=table value=2d_gels>2-D gels</td><td rowspan=2><input type=radio name=action value=add>ADD</td>
      <tr><td><input type=radio name=table value=domain_motif>domain and
motif</td></tr>
      <tr><td><input type=radio name=table value=total_identified_proteins>total
identified proteins</td><td rowspan=2><input type=radio name=action value=delete>DELETE</td>
      <tr><td><input type=radio name=table value=mutants>mutants</td><td rowspan=2><input type=radio name=action value=update>UPDATE</td>
      <tr><td><input type=radio name=table value=protein_spots>protein
spots</td><td><input type=submit value=Submit'></td>
      </form>
FORM

print "</table>";

} else {
    print "The wrong password!\n";
    print $cgi->a({-href=> "login_interface.pl"}, "Login again");
}
else {
    print "You are not signed up!\n";
}

} else {
    if ($cgi->param("table") eq "2d_gels") {
        if ($cgi->param("action") eq "add") {
            add_2d_gels();
        }
        if ($cgi->param("action") eq "delete") {
            delete_2d_gels();
        }
        if ($cgi->param("action") eq "update") {
            update_2d_gels();
        }
    }
    if ($cgi->param("table") eq "domain_motif") {
        if ($cgi->param("action") eq "add") {
            add_domain_motif();
        }
        if ($cgi->param("action") eq "delete") {
            delete_domain_motif();
        }
        if ($cgi->param("action") eq "update") {
            update_domain_motif();
        }
    }
    if ($cgi->param("table") eq "total_identified_proteins") {

if ($cgi->param("action") eq "add")
{
    add_total();
}
if ($cgi->param("action") eq "delete")
{
    delete_total();
}
if ($cgi->param("action") eq "update")
{
    update_total();
}

if ($cgi->param("table") eq "mutants")
{
    if ($cgi->param("action") eq "add")
    {
        add_mutants();
    }
    if ($cgi->param("action") eq "delete")
    {
        delete_mutants();
    }
    if ($cgi->param("action") eq "update")
    {
        update_mutants();
    }
}

if ($cgi->param("table") eq "protein_spots")
{
    if ($cgi->param("action") eq "add")
    {
        add_protein_spots();
    }
    if ($cgi->param("action") eq "delete")
    {
        delete_protein_spots();
    }
    if ($cgi->param("action") eq "update")
    {
        update_protein_spots();
    }
}

print "</pre>";
print $cgi->end_html();

#Add a record to the table total_identified_proteins
sub add_total
{
    print "<font color=blue>Please input the following information:</font>
;
    print <<TABLE;
    <table border=0 width=50% bgcolor='beige'>
        <form action="add_tables.pl">
            <tr><td bgcolor='yellow' colspan='2' align='center'><input type='radio' name='table' value='total_identified_proteins' checked> Add a record to the table total_identified_proteins</td></tr>
            <tr><th align='left'>InputID</th><td><input type='text' name='inputID'></td></tr>
            <tr><th align='left'>Protein Name</th><td><input type='text' name='name'></td></tr>
            <tr><th align='left'>Deduced mass</th><td><input type='text' name='mass'></td></tr>
            <tr><th align='left'>Deduced pl</th><td><input type='text' name='pl'></td></tr>
            <tr><th align='left'>Cyanobase No.</th><td><input type='text' name='cyano_ID'></td></tr>
            <tr><th align='left'>NCBI No.</th><td><input type='text' name='NCBI_ID'></td></tr>
            <tr><td align='center' colspan='2' x<input type='submit' value='Submit'></td></tr>
        </form>
    </table>
    TABLE
}

sub add_2d_gels()
{
    print "Please add the following information:
;
    print <<TABLE;
    <table border=0 width=50% bgcolor='beige'>
        <form action="add_tables.pl">
            <tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='2d_gels' checked> Add a record to the table 2d_gels</th></tr>
            <tr><th align='left'>Gel ID</th><td><input type='text' name='Gel_ID'></td></tr>
            <tr><th align='left'>Description</th><td><input type='text' name='Description'></td></tr>
            <tr><td align='center' colspan='2' x<input type='submit' value='Submit'></td></tr>
        </form>
    </table>
    TABLE
}

sub add_domain_motif()
{
    print "Please add the following information:
;
    print <<TABLE;
    <table border=0 width=50% bgcolor='beige'>
        <form action="add_tables.pl">
            <tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='domain_motif' checked> Add a record to the table domain_motif</th></tr>
            <tr><th align='left'>Domain name</th><td><input type='text' name='D_Name'></td></tr>
            <tr><th align='left'>Protein ID</th><td><input type='text' name='P_ID'></td></tr>
            <tr><td align='center' colspan='2' x<input type='submit' value='Submit'></td></tr>
        </form>
    </table>
    TABLE
sub add_mutants()
{
  print "Please add the following information:\n";
  print <<TABLE;
  <table border=0 width=50% bgcolor='beige'>
  <form action='add_tables.pl'>
  <tr><th bgcolor='red' colspan=2 align='center'><input type='radio' name='table' value='mutants' checked>Add a record to the table mutants</th></tr>
  <tr><th align='left'>Protein ID</th><td><input type='text' name='P_ID'></td></tr>
  <tr><th align='left'>Protein Name</th><td><input type='text' name='P_Name'></td></tr>
  <tr><th align='left'>Investigator</th><td><input type='text' name='Investigator'></td></tr>
  <tr><th align='left'>Phenotype</th><td><input type='text' name='Phenotype'></td></tr>
  <tr><th colspan=2 align='center'><input type='submit' value='Submit'></th></tr>
  </form>
</table>
TABLE
}

sub add_protein_spots()
{
  print "Please add the following information:\n";
  print <<TABLE;
  <table border=0 width=50% bgcolor='beige'>
  <form action='add_tables.pl'>
  <tr><th bgcolor='red' colspan=2 align='center'><input type='radio' name='table' value='protein_spots' checked>Add a record to the table protein_spots</th></tr>
  <tr><th align='left'>Spot ID</th><td><input type='text' name='Spot_ID'></td></tr>
  <tr><th align='left'>Gel ID</th><td><input type='text' name='Gel_ID'></td></tr>
  <tr><th align='left'>Protein ID</th><td><input type='text' name='cyano_ID'></td></tr>
  <tr><th colspan=2 align='center'><input type='submit' value='Submit'></th></tr>
  </form>
</table>
TABLE
}

sub delete_total()
{
  print "<body bgcolor='beige'>";
  print $cgi->h2("Please input one of the following for deletion: <br>");
  print <<TABLE;
  <table border=1 width=100%>
  <form action='delete_table.pl'>
  <tr><th bgcolor='red' colspan=2 align='center'><input type='radio' name='table' value='total_identified_proteins' checked>Delete a record from the table total_identified_proteins</th></tr>
  <tr><td>Input No.</td><td><input type='text' name='inputID'></td></tr>
  <tr><td>Cyanobase No.</td><td><input type='text' name='cyano_ID'></td></tr>
  <tr><td>NCBI No.</td><td><input type='text' name='NCBI_ID'></td></tr>
  <tr><td colspan=2 align='center'><input type='submit' value='Submit'></td></tr>
  </form>
</TABLE

print "</body>";
sub delete_2d_gels
{
    print "<body bgcolor='beige'>";
    print $cgi->h2("Please input one of the following for deletion: <br>");
    print "<form action="delete_table.pl">
    <tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='2d_gels' checked>Delete a record from the table 2d_gels</th></tr>
    <tr><td>Gel_ID</td><td><input type='text' name='Gel_ID'></td></tr>
    <tr><td colspan='2' align='center'><input type='submit' value='Submit'></td></tr>
</form>
</table>

TABLE
print "</body>";
}

sub delete_domain_motif
{
    print "<body bgcolor='beige'>";
    print $cgi->h2("Please input one of the following for deletion: <br>");
    print "<form action="delete_table.pl">
    <tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='domain_motif' checked>Delete a record from the table domain_motif</th></tr>
    <tr><td>Protein_ID</td><td><input type='text' name='P_ID'></td></tr>
    <tr><td colspan='2' align='center'><input type='submit' value='Submit'></td></tr>
</form>
</table>

TABLE
print "</body>";
}

sub delete_protein_spots
{
    print "<body bgcolor='beige'>";
    print $cgi->h2("Please fill all the forms for deletion: <br>");
    print "<form action="delete_table.pl">
    <tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='protein_spots' checked>Delete a record from the table protein_spots</th></tr>
    <tr><td>Spot ID</td><td><input type='text' name='Spot_ID'></td></tr>
    <tr><td>Gel ID</td><td><input type='text' name='Gel_ID'></td></tr>
    <tr><td colspan='2' align='center'><input type='submit' value='Submit'></td></tr>
</form>
</table>

TABLE
print "</body>";
sub delete_mutants
{
    print "\n<body bgcolor=beige>\n";
    print $cgi->h2("Please input one of the following for deletion: <br>");
    print "\n\n<table border=1 width=100%>\n<form action="delete_table.pl">\n<tr><th bgcolor='red' colspan='2' align='center'><input type='radio' name='table' value='mutants' checked>Delete a record from the table mutants</th></tr>
<tr><td>Protein_ID</td><td><input type="text" name="P_ID"></td></tr>
<tr><td colspan='2' align='center'><input type='submit' value='Submit'></td></tr>
</form>
</table>\nTABLE\nprint "</body>";
}

sub updatetotal
{
    print "<font color=blue>Updatet the table total_identified_proteins:</font>\n";
    print "\n\n<table border=0 width=50% bgcolor='beige'>\n<form action="update_table.pl">\n<tr><td bgcolor='yellow' colspan='2' align='center'><input type='radio' name='table' value='total_identified_proteins' checked>Update a record in the table total_identified_proteins</td></tr>
<tr><th align='left'>InputID</th><td><input type='text' name='inputID'></td></tr>
<tr><th align='left'>Protein Name</th><td><input type='text' name='name'></td></tr>
<tr><th align='left'>Deduced mass</th><td><input type='text' name='mass'></td></tr>
<tr><th align='left'>Deduced pl</th><td><input type='text' name='pl'></td></tr>
<tr><th align='left'>Cyanobase No.</th><td><input type='text' name='cyano_ID'></td></tr>
<tr><th align='left'>NCBI No.</th><td><input type='text' name='NCBI_ID'></td></tr>
<tr><td align='center' colspan='2'><input type='submit' value='Submit'></td></tr>
</form>
</table>\nTABLE\n}

sub update_2d_gels
{
    print "<font color=blue>Updatet the table 2d_gels:</font>\n";
    print "\n\n<table border=0 width=50% bgcolor='beige'>\n<form action="update_table.pl">\n<tr><td bgcolor='yellow' colspan='2' align='center'><input type='radio' name='table' value='2d_gels' checked>Update a record in the table 2d_gels</td></tr>
<tr><th align='left'>Gel_ID</th><td><input type='text' name='Gel_ID'></td></tr>
<tr><th align='left'>Description</th><td><input type='text' name='Description'></td></tr>
<tr><td align='center' colspan='2'><input type='submit' value='Submit'></td></tr>
</form>
</table>\nTABLE\n}";
sub update_protein_spots
{print $cgi->h3("No update action allowed for this table!");}

sub update_mutants
{
    print "<font color=blue>Update the table mutants:</font>
    <form action="update_table.pl">
    <tr><tdbgcolor='yellow' colspan='2' align='center'><input type='radio' name='table' value='mutants' checked>
    <input type='submit' value='Submit'></td></tr>
    </form>
    </table>
    
sub updatedomainmotif
{print $cgi->h3("No update action allowed for this table!");}

C5. The source code for the program compare.pl

#!/c:/perl\bin\perl
#
#The program was developed to compare identified proteins between different research groups.
#Users can either input the file name containing ID of identified proteins or directly
#input the ID of identified proteins to compare with the identified proteins in the database.
#
use CGI;
use DBI;
use strict;
use FileHandle;

my ($cgi)= new CGI;
my ($dsn)= "DBI:mysql:cyano_membrane_protein:localhost";
my ($user_name)= undef;
my ($password)= undef;
my ($dbh, $sth);
my ($file)=$cgi->param("file_name");
my ($access_nums)=$cgi->param("access_nums");
my (@line);
my (@ary2);#used to hold the Cyanobase accession numbers.
my (@ary2);#hold the column name of the table.
my ($num_of_matched)=0;

print $cgi->header ();
print $cgi->start_html (-title => "Compare the identified proteins with the database");
print "<pre>";
$dbh=DBI->connect($dsn, $user_name, $password, {RaiseError=>1}); #connect to database.

if($file ne "")
{

# Open the file to get accession numbers.
while (<$file>){chomp($_); push(@line, $_);} foreach my $elm (@line)
{
    my (@ary)= split('t', $elm);
    push(@nums, $ary[0]);
    @ary=();
}
if ($access_nums ne "")
{
    @nums=split('n', $access_nums);
    for (my $j=0;$j<scalar(@nums);$j++)
    {
        if ($j==(scalar(@nums)-1))
        {
            my ($last_elms)=substr($nums[$j], 0, 7);
            pop(@nums);
            push(@nums, $last_elms);
        }
        else
            {chop($nums[$j]);}# It is a trick, chomp does not work here.
    }
}

# Get cyanobase no. from the array @nums one by one and search against the database, print
# the search result in a table
print "<TABLE BORDER>\n";
for (my $i=0;$i<scalar(@nums);$i++)
{
    my ($number)=$nums[$i];
    $sth=$dbh->prepare(qq{
        select p.inputID, p.name, p.cyano_ID, p.NCBI_ID, p.mass, p.pl, t.gravay, t.TM
        from total_identified_proteins p, tmhelix t
        where (p.cyano_ID='$number') and (p.cyano_ID=t.proteinID) });
    my $sth1=$dbh->prepare(qq{
        select count(*)
        from total_identified_proteins p
        where (p.cyano_ID='$number') });
    $sth->execute();
    $sth1->execute();
    if ($i==0)
    {
        display_table_header();# Print the name of columns.
    }
    my(@ary_for_count) = $sth1->fetchrow_array();
    my ($count)=$ary_for_count[0];
    if ($count!=0)


```perl
{ display_table_rows(); #Print contents of each row
  $num_of_matched++;
}
$sth->finish();
$sth1->finish();

#-----------------------------------
print "</TABLE BORDER>n"
if ($num_of_matched==0)
{
  print "No protein was identified here! <br>";
}
else
{
  print "<br><font color=blue>$num_of_matched proteins were identified in the database.</font> <br>";
}
$dbh->disconnect();
print "</pre>";
print $cgi->end_html();
#end of main program

#----------------------------------

sub display_table_header
{
  print "<TR>n"
  foreach my $col_name1 (@{$sth->{NAME}})
  {
    display_cell ("TH", $col_name1, 0);
  }
  print "</TR>n"
}

sub display_table_rows
{
  foreach my $col_name (@{$sth->{NAME}})
  {
    push(@ary2, $col_name);
  }

  # display table rows
  while (my @ary3 = $sth->fetchrow_array ())
  {
    print "<TR>n",
    for ( my $i=0;$i<scalar(@ary3);$i++)
    {
      if ($ary2[$i] eq "ID_in_NCBI" || $ary2[$i] eq "NCBI_ID")
      {
        $url2.="&db=protein&list_uids=$ary3[$i]&dopt=GenPept";
        display_cell("TD", $cgi->a( {-href => $url2}, $ary3[$i]), 0);
      }
      elsif ($ary2[$i] eq "proteinID" || $ary2[$i] eq "PID" || $ary2[$i] eq "cyano_ID")
      {
        print "$ary2[$i]"
      }
      else
      {
```
sub display_cell
{
    my ($tag, $value, $encode) = @_;
    $value = $cgi->escapeHTML ($value) if $encode;
    $value = "0" unless $value;
    print "<$tag>$value</$tag>\n";
}

C6. The source code for the program add_table.pl

#!/c:\perl\bin\perl
# This program was developed to add data to the database. Users can input data to each table of the database from the provided interface.
# use CGI;
use DBI;
use strict;

my ($cgi)= new CGI;
my ($dsn)= "DBI:mysql:cyano_membrane_protein:localhost";
my ($user_name)= undef;
my ($password)= undef;
my ($dbh, $sth);
my ($rows);
my ($tb_name)=$cgi->param("table");
print $cgi->header ();
print $cgi->start_html (-title => "Add record to the peripheral proetins resolved on 2-D gel with pH 3-10");
#connect to database
$dbh=DBI->connect($dsn, $user_name, $password, {RaiseError=>1});

#The following code try to add data to different tables.
if ($tb_name eq "total_identified_proteins")
{
    my ($inputID)= $cgi->param("inputID");
    my ($name)= $cgi->param("name");
    my ($cyano_ID)= $cgi->param("cyano_ID");
my ($NCBI_ID) = $cgi->param("NCBI_ID");
my ($mass) = $cgi->param("mass");
my ($pl) = $cgi->param("pl");
$sth = $dbh->prepare(qq {INSERT $tb_name(inputID, name, mass, pl, cyano_ID, NCBI_ID)
VALUES('$inputID', '$name', '$mass', '$pl', '$cyano_ID', '$NCBI_ID')});

if ($tb_name eq "2d_gels")
{
    my ($Gel_ID) = $cgi->param("Gel_ID");
    my ($Description) = $cgi->param("Description");
    $sth = $dbh->prepare(qq {INSERT $tb_name(Gel_ID, Description)
VALUES('$Gel_ID', '$Description')});
}

if ($tb_name eq "domain_motifs")
{
    my ($D_Name) = $cgi->param("D_Name");
    my ($P_ID) = $cgi->param("P_ID");
    $sth = $dbh->prepare(qq {INSERT $tb_name(D_Name, P_ID)
VALUES('$D_Name', '$P_ID')});
}

if ($tb_name eq "mutants")
{
    my ($P_ID) = $cgi->param("P_ID");
    my ($P_Name) = $cgi->param("P_Name");
    my ($Investigator) = $cgi->param("Investigator");
    my ($Phenotype) = $cgi->param("Phenotype");
    $sth = $dbh->prepare(qq {INSERT $tb_name(P_ID, P_Name, Investigator, Phenotype)
VALUES('$P_ID', '$P_Name', '$Investigator', '$Phenotype')});
}

if ($tb_name eq "protein_spots")
{
    my ($Spot_ID) = $cgi->param("Spot_ID");
    my ($Gel_ID) = $cgi->param("Gel_ID");
    my ($cyano_ID) = $cgi->param("cyano_ID");
    $sth = $dbh->prepare(qq {INSERT $tb_name(Spot_ID, Gel_ID, cyano_ID)
VALUES('$Spot_ID', '$Gel_ID', '$cyano_ID')});
}

$rows = $sth->execute();
$sth->finish();
$dbh->disconnect();

print $cgi->h1("The record have been added to the database!

C7. The source code for the program delete.pl

#! c:\perl\bin\perl
use CGI;
use DBI;
use strict;

my ($cgi)= new CGI;
my ($dsn)= "DBI:mysql:cyano_membrane_protein:localhost";
my ($user_name)= undef;
my ($password)=undef;
my ($dbh, $sth);
my ($tb_name)=$cgi->param("table");
my ($rows);

print $cgi->header();
print $cgi->start_html (-title => "Delete record from table total_identified_proteins");
print "<pre>";

#connect to database
$dbh=DBI->connect($dsn, $user_name, $password, {RaiseError=>1});

#Delete a record from corresponding tables.
if ($tb_name eq "total_identified_proteins")
{
    my ($inputID)=$cgi->param("inputID");
    my ($cyano_ID)=$cgi->param("cyano_ID");
    my ($NCBI_ID)=$cgi->param("NCBI_ID");
    if ($cyano_ID eq "" && $NCBI_ID eq "")
    {
        $sth=$dbh->prepare("delete from total_identified_proteins where inputID='$inputID'");
    }
    if ($inputID eq "" && $NCBI_ID eq "")
    {
        $sth=$dbh->prepare("delete from total_identified_proteins where cyano_ID='$cyano_ID'");
    }
    if ($inputID eq "" && $cyano_ID eq "")
    {
        $sth=$dbh->prepare("delete from total_identified_proteins where NCBI_ID='$NCBI_ID'";
    }
}

if ($tb_name eq "2d_gels")
{
    my ($Gel_ID)=$cgi->param("Gel_ID");
    $sth=$dbh->prepare("delete from $tb_name where Gel_ID='$Gel_ID'";
}

if ($tb_name eq "domain_motif")
{
    my ($P_ID)=$cgi->param("P_ID");
    $sth=$dbh->prepare("delete from $tb_name where P_ID='$P_ID'";
}

if ($tb_name eq "protein_spots")
{
my ($Gel_ID)=$cgi->param("Gel_ID");
my ($Spot_ID)=$cgi->param("Spot_ID");
$sth=$dbh->prepare("delete from $tb_name where Gel_ID='$Gel_ID' and Spot_ID='$Spot_ID'");
}
if ($tb_name eq "mutants")
{
    my ($P_ID)=$cgi->param("P_ID");
    $sth=$dbh->prepare("delete from $tb_name where P_ID='$P_ID'");
}
$sth->execute();
$sth->finish();
$dbh->disconnect();
print "The record have been deleted from the database!\n";
print "</pre>";
print $cgi->end_html();
exit(0);

C8. The source code for the program update_table.pl

#! c:\perl\bin\perl
#This program is developed to change the data of an existing record in the database.
use CGI;
use DBI;
use strict;
my ($cgi) = new CGI;
my ($dsn) = "DBI:mysql:cyano_membrane_protein:localhost";
my ($user_name) = undef;
my ($password) = undef;
my ($dbh, $sth);
my ($rows);
my ($inputID) = $cgi->param("inputID");
my ($name) = $cgi->param("name");
my ($mass) = $cgi->param("mass");
my ($pl) = $cgi->param("pl");
my ($cyano_ID) = $cgi->param("cyano_ID");
my ($NCBI_ID) = $cgi->param("NCBI_ID");
my ($P_ID) = $cgi->param("P_ID");
my ($Gel_ID) = $cgi->param("Gel_ID");
my ($Investigator) = $cgi->param("Investigator");
my ($Phenotype) = $cgi->param("Phenotype");
my ($Description) = $cgi->param("Description");
my ($tb_name) = $cgi->param("table");

#connect to database
$dbh=DBI->connect($dsn, $user_name, $password, {RaiseError=>1});

#Change the data in the corresponding tables.
if ($tb_name eq "total_identified_proteins")
{

$sth=$dbh->prepare(qq{
    UPDATE total_identified_proteins
    SET name='$name', mass='$mass', pl='$pl', cyano_ID='$cyano_ID', NCBI_ID='$NCBI_ID'
    where inputID='$inputID'});
$rows=$sth->execute();

if ($tb_name eq "mutants")
{
    $sth=$dbh->prepare(qq{
        UPDATE mutants
        SET Investigator='$Investigator', Phenotype='$Phenotype'
        where P_ID='$P_ID'});
    $rows=$sth->execute();
}

if ($tb_name eq "2d_gels")
{
    $sth=$dbh->prepare(qq{
        UPDATE mutants
        SET Description='$Description'
        where Gel_ID='$Gel_ID'});
    $rows=$sth->execute();
}

$sth->finish();
$dbh->disconnect();

print $cgi->header();
print $cgi->start_html (-title => "Update the tables in the membrane proteome database");
print $cgi->h1("The query result:");
print "<pre>",
print $cgi->h1("The record have been updated!");
print "</pre>",
print $cgi->end_html();
exit(0);
ACKNOWLEDGEMENTS

To get a Ph.D. degree is a dream I have since I was in undergraduate college, and this dream is coming true after five years of study and research in Iowa State University. I would like to thank all of those who helped me on my way towards the doctorate degree.

First of all, I want to thank my major professor, Dr. Parag R. Chitnis. I would never forget the chance he gave me to choose a research project I was interested in, and help me to conduct research in a field that I was not familiar with. He was so patient and knowledgeable, and I will always remember the time that he explained to me everything I did not understand and help me to design my experiments. As an international student, I am extremely thankful to him for the help to solve the problem I have with my family, I would never forget that it was him who helped my wife to get a visa to join me by writing letters, making calls to the US embassy again and again. He is a great person, both as a scientist or as an advisor, I am really lucky to have him as my major professor.

Second, I want to thank my co-major professor Dr. Suresh Kothari and the rest of my POS committee members: Dr. Steven R. Rodermel, Dr. Basil J. Nikolau, Dr. Alan M. Myers and Dr. Xiaoqiu Huang. Dr. Kothari have helped me a lot in process automation using computer programs, the programs we developed together have been proven valuable in my research. Dr. Rodermel, Dr. Nikolau and Dr. Huang attend all of the annual meetings and gave me a lot of valuable suggestions in designing experiments, taking classes. Dr. Myers was so kind to serve as my committee member when Dr. Nikolau cannot present in my final oral exam.

I am also very thankful to the members of Chitnis lab. Wu was my classmates in China, I would never forget the day he picked me up from the airport and help me to settle down when I came to AMES for the first time, he was always ready to help whenever I got problems. Jun instructed me to do experiments for the first six months in the lab. Don helped me to correct my preliminary exam proposal and gave me a lot of valuable suggestions. Wade taught me a lot in equipment analysis. Mike is a happy boy and always willing to help. I would also thank all the other members who helped me in chitnis lab: Sushma, Laysa, Jianmin, Huadong, Xiaotao, Maria, Dondra and in kothari lab: Donald and Kent.
I visited Dr. Donald A. Bryant’s Lab in PSU last winter. Dr. Gaozhong Sheng helped me to conduct the fluorescence induction experiment and Dr. Bryant’s gave me a lot of suggestions, we got a many interesting results during my visit. Thank them all for their help and effort.

I would also thank Trish, a secretary in the department of Biochemistry, Biophysics and Molecular Biology, she is always helpful, for my paper work, for my and my wife’s visa. I could not even remember how much help I got from her. Linda and Kathy are secretaries of my major program genetics and BCB, I would thank them for their help during my five-year study.

My wife, Wei Wang is always in support with me, even in the two years she stayed alone in China. She will share my happiness of being a doctor of philosophy. I want to thank my parents, my brothers and sisters and my wife’s parents, their support always help me to go forward.