# Probability-Based Scoring Function as a Software Tool Used in the Genome-Based Identification of Proteins from *Spirulina platensis*

Wimada Thammasorn[1], Korakot Eadjongdee[1], Apiradee Hongsthong*[,2], Kriengkrai Porkaew[1] and Supapon Cheevadhanarak[3]

[1]*School of Information Technology, King Mongkut's University of Technology Thonburi, 126 Pracha-U-Thit Rd., Bangmod, Thungkru, Bangkok 10140, Thailand;* [2]*BEC Unit, KMUTT-Bangkhuntien, 83 Moo 8, Thakham, Bangkhuntien, Bangkok 10150, Thailand;* [3]*School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand*

**Abstract:** One of the major goals of proteomic research is the identification of proteins, a goal that often requires various software tools and databases. These tools have to be able to handle large amounts of data, such as those generated by PMF (Peptide Mass Fingerprinting), a high throughput technique. A newly sequenced organism, *Spirulina platensis*, was recently used to generate an *in silico* database, and thus an in-house tool designed for compatibility with this database and its inputs (PMF) was constructed in the present study. With a probability based scoring function, this tool effectively ranked ambiguous protein identification results by using five criteria: score, number of matched peptides, % coverage, pI and molecular weight. As a result, the protein identification step of *Spirulina* proteomic studies can be achieved precisely. Moreover, a very useful function of this tool is its capability for batch processing, in which the system can handle protein-identification searches of a hundred of proteins automatically, from a single user's input. Therefore, the tool not only gives accurate protein identification results but also saves the user time in processing a large amount of data.

**Keywords:** Peptide Mass Fingerprinting (PMF), *S. platensis*, 2D-DIGE, Protein isoelectric points (pI), Bisection method and Probability-based scoring function.

## INTRODUCTION

The identification of differentially expressed proteins by employing experimental techniques and tools is a goal of proteomics. After protein isolation by various techniques, including two dimensional- gel electrophoresis (2D-PAGE), a protein of interest is subjected to protein identification by using mass spectrometry techniques coupled with bioinformatics. Mass spectrometry (MS) technology is widely used to identify proteins by generating either the peptide mass fingerprinting (PMF) or the peptide fragmentation fingerprinting (PFF) of proteins of interest. Then, these PMFs and PFFs are searched against PMFs and PFFs in available databases, to identify the proteins. PMFs are analyzed by comparing an experimental mass list with theoretical mass lists in databases. This identification step requires both efficient software tools and appropriate databases to obtain reliable protein identification results.

At present, several software tools for protein identification using PMF have been constructed, including MASCOT, the MS-Fit tool, the ALDENTE tool, etc. These tools use several algorithms to calculate the scores of matched proteins, e.g. the Bayesian theory of probability-based scoring methods, the genetic algorithm method [1] and HMMs (Hidden Markov models) [2]. However, the tool constructed in the present study was designed not only to serve the needs of users but also to assist in a precise and time-saving protein identification process by accurate scoring plus ranking function, pI-filtering and PMF-batch processing. Therefore, an *in silico* database and an in-house software tool were constructed for *S. platensis* protein identification by using PMFs as inputs. In our previous study, simple ranking methods were employed, by counting the number of matched peptides and calculating the coverage percentage of the matched peptides compared to the whole proteins, in order to rank ambiguous protein identification results. However, an effective and accurate scoring method is required to get rid of ambiguous data and make protein identification more reliable. Thus, in the present study, an effective scoring method was developed using a probability-based scoring function. Moreover, the isoelectric point (pI) and molecular weight values of a protein were also used as criteria to pick out a target protein among redundant protein identification results, which may contain proteins with very close scores. For effective use of the tool, a batch processing module was also developed to handle hundreds of inputs simultaneously in a single run.

## MATERIALS AND METHODS

### Programming Software

In this study, the PHP programming language was used to write code to calculate protein isoelectric point values and develop a probability-based scoring function, including a connection to the database. In the case of this tool, Apache 2.5.10 was employed to view all data from a database, which was installed with Navicat MySQL version 5.0.5. Moreover,

*Address correspondence to this author at the BEC Unit, KMUTT, 83 Moo8, Thakham, Bangkhuntien, Bangkok 10150, Thailand; Tel: 662-470-7509; Fax: 662-452-3455;
E-mails: apiradee@biotec.or.th, apiradee@pdti.kmutt.ac.th

1875-0362/09 

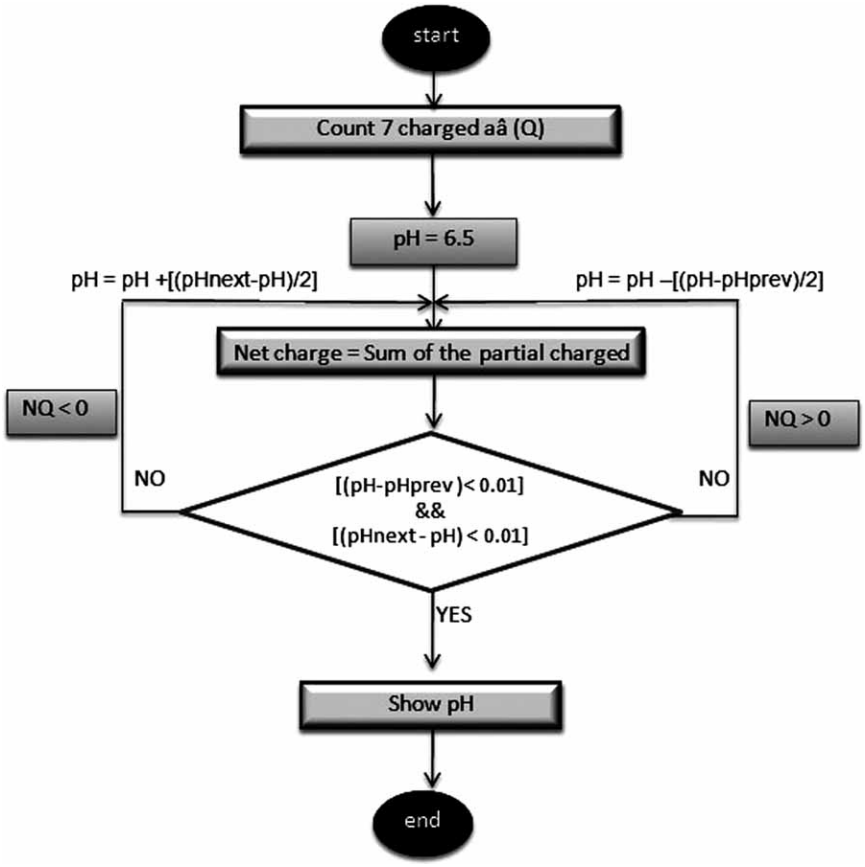



**Fig. (1).** Bisection method used for pI calculation.

phpMyAdmin version 2.10.3 was used to manage the database and perform tasks such as creating tables. To manage the web interface, the input and output interfaces of this tool were created as Graphic User Interfaces (GUIs) in the HTML language employed in the Macromedia Dreamweaver MX program.

### pI Calculation Method

The pI of an individual protein was calculated by employing the Bisection method [3, 4] as shown in Fig. (**1**). The pH values, which are related to the total charge of proteins, were divided into two sections, pH 0-7 and 8-14, in order to determine total charge. If the total charge of a protein is near zero at a certain pH value, then the pH value at this state will be the pI. However, the total charge of each protein was calculated from the summation of the charges of its constituent amino acids. These macromolecules can be categorized into two major groups: (*i*) the group of positively charged amino acids, which consists of histidine (H), lysine (K) and arginine (R), and (*ii*) the group of negatively charged amino acids, which consists of aspartate (D), glutamate (E), cysteine (C) and tyrosine (Y). In order to reduce the time required for the pI calculation, the pKs values of the seven charged amino acids [4] were collected in a MySQL database. Then, the PHP programming language was used to calculate the pI values of an individual protein from the *S. platensis* database. Finally, the theoretical pI was shown in the 'Hits' protein results on the GUI.

### Probability–Based Scoring Function Method

The scores of peptides that matched theoretical peptides from the *Spirulina*-PMF database were calculated in the form of probabilities. In the first step, the total charge of each protein was obtained from the summation of the negative and positive charges of each macromolecule, as shown in equations 1 and 2 [4], where $pK_{Ni}$ and $pK_{pi}$ represent the pK of negative and positive macromolecules of size i.

Charge of negative macromolecules:

$$\sum_{i}^{n}=1\frac{-1}{1+10^{\mathrm{pk}}\,\mathrm{Ni}^{-\mathrm{pH}}} \qquad \text{Equation (1)}$$

Charge of positive macromolecules:

$$\sum_{i}^{n}=1\frac{-1}{1+10^{\mathrm{pH\text{-}pK}}\,\mathrm{Pi}} \qquad \text{Equation (2)}$$

In the case of the probability-based scoring function, the input peptides were matched with proteins in the database, and probabilities and scores were calculated, as shown in equation 3 and 4, respectively, where $\alpha$ represents a constant between zero and one, $m_{ij}$ is the number of the matched peptide, and $M_j$ is the number of all peptides obtained from digestion of the theoretical protein.

$$\Pr(P_k)=\prod_{i=R(l),l\in H_k}\left(1-\left(\frac{m_{ij}}{M_j}(1-\alpha)+\alpha\right)^{n_{ij}}\right)^{k} \qquad \text{Equation (3)}$$

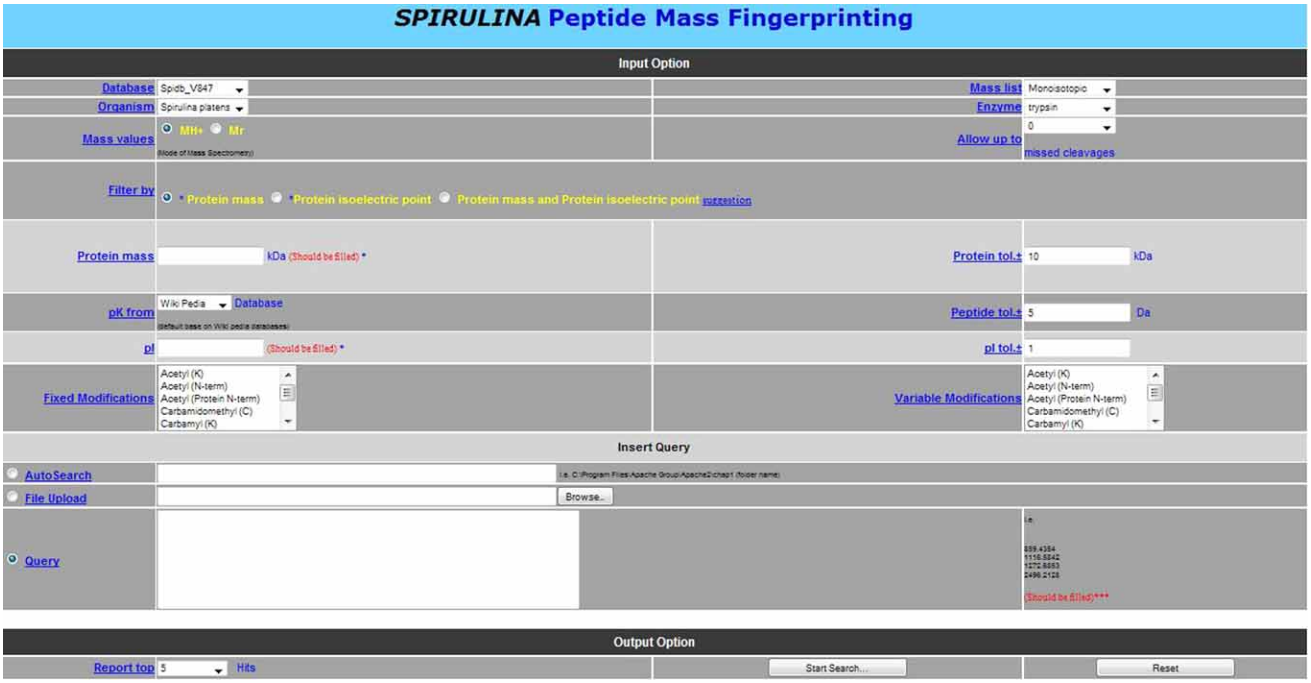



**Fig. (2).** Input interface of *Spirulina* database - PMF software tool.

$$\text{Score} = -\log \Pr(Pk) \qquad \text{Equation (4)}$$

If the input-peptide does not match with a theoretical-peptide in the database, the probability of this input-peptide is zero. Thus, the score has values between zero and one, which requires visualizing the results in the form of decimal numbers. For ease of comparison, these probabilities are converted into logarithmic form. For example, if the probabilities are 0.0001 and 0.002, these scores in logarithmic form will be -4 and -3.6989 which are simpler for users to analyze. Finally, the scores of the 'Hits' protein results are shown on the GUI.

### Tool Validation

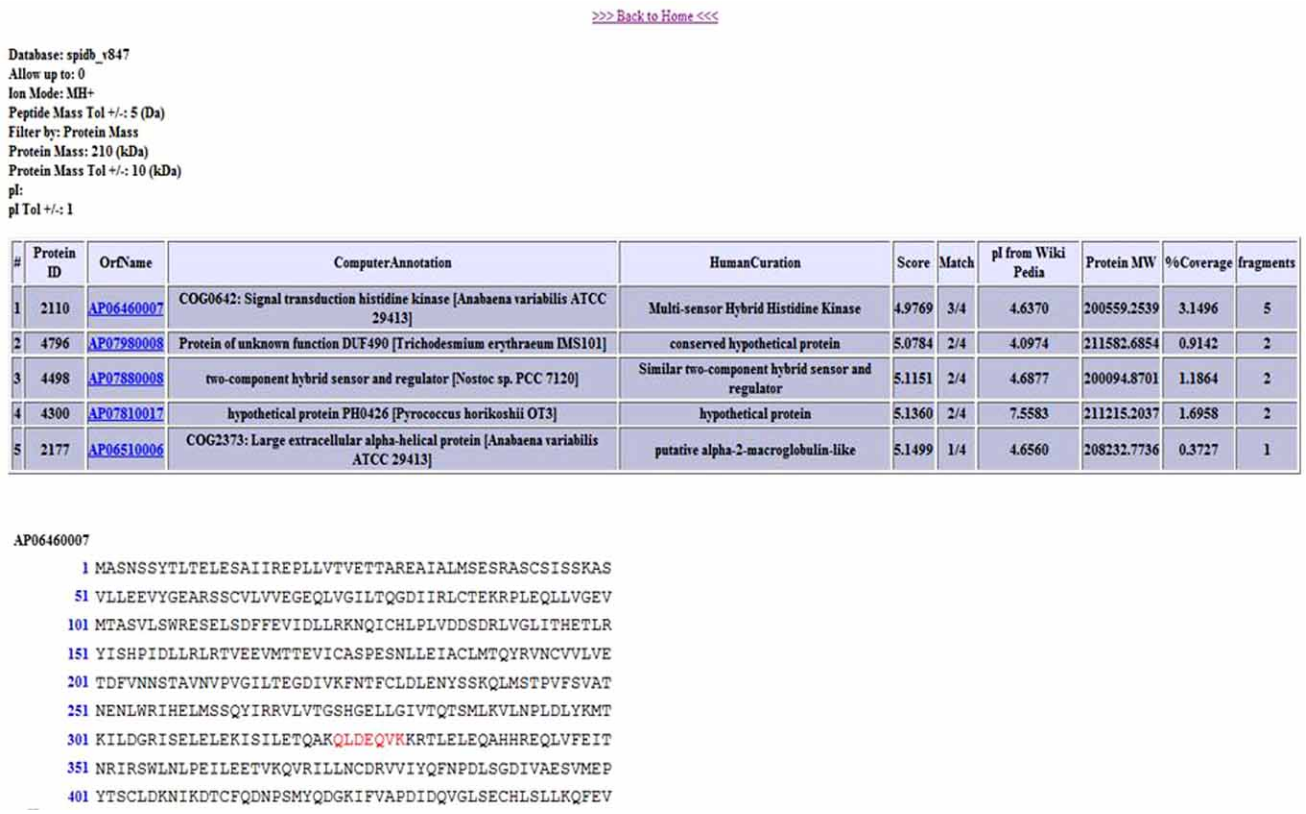
For tool validation, cross-species proteins were used to check an accuracy of the current tool, in order to compare the 'Hits' results with online PMF tools such as the Mascot tool, by using two main methods. First, cross-species proteins such as the photosystem II D1 protein found in cyanobactria (*Synechocystis sp.PCC6803* (*slr0752*)), and the pyruvate kinase of *Escherichia coli K-12 MG1655* were digested, in order to collect their PMF data using miss cleavage values of zero to three. Second, the PMF data of known proteins, obtained from the 2-DE technique in the NCBI database (http://www.ncbi.nlm.nih.gov/Ftp/) were searched for with the tool. Finally, the *E. coli* genome was downloaded to the database and then PMFs of the known proteins of *E. coli* were searched with this tool.

## RESULTS & DISCUSSION

The input interface of the current version of the in-house software tool was designed as shown in Fig. (**2**). By using this GUI, users could fill experimental PMF data into 'Query' and 'Autosearch' sections on the input interface (Fig. **2**), together with protein mass and pI obtained from 2-DE experiments [5]. The source code of the tool is available at http://spirulina.biotec.or.th/~spirulina/proteome/index.htm.

Moreover, users could select a pI-database to represent the "Hits" protein results from EMBOSS, DTASelect, Solomon, Sillero, Rodwell and Wikipedia databases as shown in the dashed box in Fig. (**2**). In order to rank the 'Hits' protein results, five criteria were considered consecutively, (*i*) probability-based score, (*ii*) number of matched proteins, (*iii*) pI, (*vi*) protein molecular mass and (*v*) % coverage. On the output interface, all results were represented in the form of tables and lists of the amino acid sequences of theoretical protein 'Hits,' which were obtained from the *Spirulina* database under the input criteria setting (Fig. **3**). Criteria for searching are presented in a table in Fig. (**3**), such as database name (only two versions, 677 and 847), allowed missed cleavage, ion mode (MH+ or Mr Modes), type of filtering, protein mass, protein tolerance, pI and pI tolerance.

A limitation of this tool underlying the search algorithm is the required filtering of the 'Hits' results by the protein mass and pI, because of differences in their distribution patterns. However, in both versions of the database, the same pattern of protein distributions were represented in Fig. (**4a**) and Fig. (**4b**), for database versions 677 and 847, respectively. The protein mass distribution of both versions showed that the most abundance protein masses were in the range of 5 kDa to 99 kDa. Thus, if a user selected a wide gap of protein tolerance for the experimental protein within this range, the execution time would be very long. Therefore, for the experimental proteins, which have protein masses within the range of 5 kDa to 99 kDa, a user should use a protein tolerance of 0.1 kDa when searching the 'Hits' protein results. On



>>> Back to Home <<<

Database: spidb_v847
Allow up to: 0
Ion Mode: MH+
Peptide Mass Tol +/-: 5 (Da)
Filter by: Protein Mass
Protein Mass: 210 (kDa)
Protein Mass Tol +/-: 10 (kDa)
pI:
pI Tol +/-: 1

| # | Protein ID | OrfName | ComputerAnnotation | HumanCuration | Score | Match | pI from Wiki Pedia | Protein MW | %Coverage | fragments |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2110 | AP06460007 | COG0642: Signal transduction histidine kinase [Anabaena variabilis ATCC 29413] | Multi-sensor Hybrid Histidine Kinase | 4.9769 | 3/4 | 4.6370 | 200559.2539 | 3.1496 | 5 |
| 2 | 4796 | AP07980008 | Protein of unknown function DUF490 [Trichodesmium erythraeum IMS101] | conserved hypothetical protein | 5.0784 | 2/4 | 4.0974 | 211582.6854 | 0.9142 | 2 |
| 3 | 4498 | AP07880008 | two-component hybrid sensor and regulator [Nostoc sp. PCC 7120] | Similar two-component hybrid sensor and regulator | 5.1151 | 2/4 | 4.6877 | 200094.8701 | 1.1864 | 2 |
| 4 | 4300 | AP07810017 | hypothetical protein PH0426 [Pyrococcus horikoshii OT3] | hypothetical protein | 5.1360 | 2/4 | 7.5583 | 211215.2037 | 1.6958 | 2 |
| 5 | 2177 | AP06510006 | COG2373: Large extracellular alpha-helical protein [Anabaena variabilis ATCC 29413] | putative alpha-2-macroglobulin-like | 5.1499 | 1/4 | 4.6560 | 208232.7736 | 0.3727 | 1 |

AP06460007

```
  1 MASNSSYTLTELESAIIREPLLVTVETTAREAIALMSESRASCSISSKAS
 51 VLLEEVYGEARSSCVLVVEGEQLVGILTQGDIIRLCTEKRPLEQLLVGEV
101 MTASVLSWRESELSDFFEVIDLLRKNQICHLPLVDDSDRLVGLITHETLR
151 YISHPIDLLRLRTVEEVMTTEVICASPESNLLEIACLMTQYRVNCVVLVE
201 TDFVNNSTAVNVPVGILTEGDIVKFNTFCLDLENYSSKQLMSTPVFSVAT
251 NENLWRIHELMSSQYIRRVLVTGSHGELLGIVTQTSMLKVLNPLDLYKMT
301 KILDGRISELELEKISILETQAKQLDEQVKKRTLELEQAHHREQLVFEIT
351 NRIRSWLNLPEILEETVKQVRILLNCDRVVIYQFNPDLSGDIVAESVMEP
401 YTSCLDKNIKDTCFQDNPSMYQDGKIFVAPDIDQVGLSECHLSLLKQFEV
```

**Fig. (3).** Output interface of *Spirulina* database – PMF software tool.

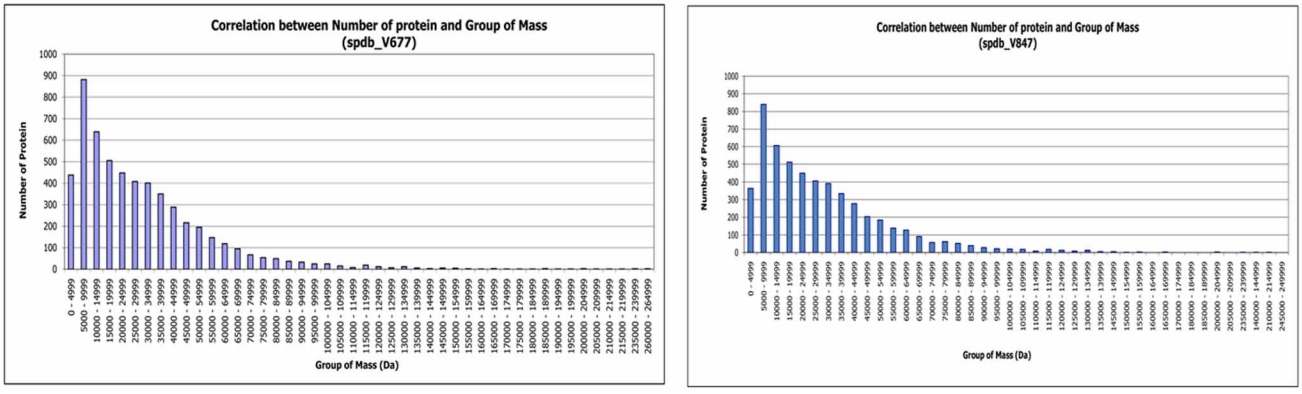



**Fig. (4).** Protein mass distribution of the *Spirulina* database: (**a**) version 677; (**b**) version 847.

the other hand, if the experimental proteins have protein masses of more than 100 kDa, a value for protein tolerance could be selected from 10 kDa to 30 kDa, resulting in a the searching time of 30 minutes.

In the case of pI distribution, the pI values of both databases differed in their pKs values from the pKs database, as shown in Fig. (**5a**) and (**b**) for versions 677 and 847, respectively. In these patterns, the most abundant pI values were found within the range of 3 to 7 and 8 to 11, for database versions 677 and 847, respectively. These ranges are very wide. Thus, if a user selected only pI filtering for protein identification, the process would take around 30 min to execute. Therefore, users are recommended to use protein mass filtering coupled with pI filtering.

For tool validation, two known proteins were searched for using the current *Spirulina* tool in order to compare these results to the 'Hits' results of Mascot. The 'Hits' results for each protein are shown in Table **1**, and Table **2** for photosystem II D1 (*Synechocystis sp.PCC6803*), and pyruvate kinase (*E. coli K-12*). The photosystem II D1 protein was found in

first place in the 'five Hits' results (Table **1**) by using the Mascot tool and our current tool. For Mascot, this protein was found with a score of 160, expected values of 1.5E-10, a matched number of eleven, a protein mass of 39.696 kDa, and coverage of 0.68% (Table **1a**). On the other hand, this protein was found in first place, using the current tool, with a score of 15.7398, three out of thirteen matched peptides, a pI of 5.386, a protein mass of 39635.9211 Da, and matched peptide coverage of 10.8635 % (Table **1b**). The input setting for Mascot was: the MSDB database (Mascot database), Taxonomy of bacteria, allowed up to one missed cleavage, and a peptide tolerance of 1.2 Da. In the current version of the tool, six different input parameters were used: the *Spirulina* database version 847, allowed zero missed cleavages, protein mass filtering at 39 kDa, a protein tolerance of 0.7 kDa, a peptide tolerance of 1.2, pI filtering of five, and a pI tolerance of one.

In the case of the other known protein used for tool validation, pyruvate kinase from *E. coli*, the search results from both tools show pyruvate kinase I second on the list. The *E. coli* database and the *Spirulina* databases were used for Mascot and the current tool, respectively, for the protein identification process. Using the Mascot tool, this protein was found in second place with a score of 62, expected values of 0.021, a match number of seven, a protein mass of 50.697 kDa, and coverage of 7% (Table **2a**). According to the current version of the PMF-*Spirulina* tool, the protein was also found in second place, with a score of 8.4036, a match number of four from seven peptide masses, a pI of 5.7986, a protein mass of 63.339 kDa, coverage of 21.19 %, and a fragment number of six, as shown in Table **2b**.

Consequently, a second set of tool validation experiments were carried out. The complete *E. coli* genomes were down-

**Table 2a. 'Hits' Results from the Mascot Tool Using PMF Data from Pyruvate Kinase I**

| Orf Name | Protein Name | Mass (Da) | Score | Expect | Matched | %Coverage |
|---|---|---|---|---|---|---|
| Q1RCS4_ECOUT | Hypothetical protein.- Escherichia coli (strain UTI89 / UPEC). | 6903 | 77 | 0.00058 | 6 | 26.00 |
| D64925 | pyruvate kinase (EC 2.7.1.40) [validated] - Escherichia coli (strain K-12) | 50697 | 62 | 0.021 | 7 | 7.00 |
| AAA24392 | ECOPK1 NID: - Escherichia coli | 50276 | 62 | 0.022 | 7 | 7.00 |
| 1A40 | phosphate-binding periplasmic protein precursor mutant A197W - Escherichia coli | 34516 | 61 | 0.024 | 6 | 9.00 |
| Q1RBC0_ECOUT | Pyruvate kinase I (EC 2.7.1.40).- Escherichia coli (strain UTI89 / UPEC). | 58665 | 57 | 0.07 | 7 | 6.00 |

**Note:** The results obtained from the Mascot software tool were identified by using five parameters for searching; PMF data are 7 fragments obtained from *in silico* digestion of pyruvate kinase I of *Escherichia coli K-12 MG1655*, Database name is MSDB (Mascot Database), Taxonomy of other bacteria, peptide tolerance at 1.2 Da, and Allowed missed cleavage up to one.

**Table 2b. 'Hits' Results from the Current PMF Tool Using PMF Data from Pyruvate Kinase I**

| # | Protein ID | Orf Name | Computer Annotation | Human Curation | Score | Match | pI | Protein MW | %Coverage | Fragments |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1183 | AP05200009 | hypothetical protein Cwat-DRAFT_4511 [Crocosphaera watsonii WH 8501] | | 8.3257 | 3/7 | 7.3899 | 63142.0551 | 2.5594 | 8 |
| 2 | 722 | AP04360002 | Pyruvate kinase [Trichodesmium erythraeum IMS101] | pyruvate kinase | 8.4036 | 4/7 | 5.7986 | 63339.6356 | 2.1959 | 6 |
| 3 | 2296 | AP06630005 | Protein of unknown function DUF6 [Trichodesmium erythraeum IMS101] | Conserve protein of unknown function DUF6 transmembrane | 8.4796 | 3/7 | 5.2400 | 62651.5798 | 1.5679 | 4 |
| 4 | 5318 | AP05840005 | Protein kinase:G-protein beta WD-40 repeat [Trichodesmium erythraeum IMS101] | putative serine/threonine kinase | 8.5911 | 2/7 | 9.4919 | 62681.8923 | 1.4388 | 5 |
| 5 | 5315 | AP05840001 | hypothetical protein Npun02001295 [Nostoc punctiforme PCC 73102] | regulatory components of sensory transduction system | 8.6070 | 2/7 | 5.9827 | 63288.5014 | 0.9042 | 4 |

**Note:** The results obtained from the previous version of the PMF tool were identified by using eight parameters for searching; PMF data are 7 fragments obtained from *in silico* digestion of pyruvate kinase I of *Escherichia coli K-12 MG1655*, Database name is SpiDB_v847 (*Spirulina* Database version 847), peptide tolerance at 1.2 Da, protein mass of 64 kDa, protein tolerance of 5 kDa, pI of 5, pI Tolerance of 1, and Allowed missed cleavage at zero.

**Table 3a. 'Hits' Results from Mascot Using PMF Data of *E. coli K-12 substr. DH10B***

| Spot No. | ORFs Name | Protein Name | Mass (Da) | Score | Expect | Matched | %Coverage |
|---|---|---|---|---|---|---|---|
| P0C0V0 | E85500 | proteinase DO (EC 3.4.21.-) precursor / heat shock protein htrA - Escherichia coli (strain O157:H7, substrain EDL933) | 49323 | 70 | 0.0031 | 7 | 22% |
| | Q1RG27_ECOUT | Periplasmic serine protease DegP (EC 3.4.21.-).- Escherichia coli (strain UTI89 / UPEC). | 49308 | 70 | 0.0031 | 7 | 22% |
| | CAA30997 | ECHTRA NID: - Escherichia coli | 51190 | 69 | 0.0039 | 7 | 21% |
| P61889 | DEECM | malate dehydrogenase (EC 1.1.1.37) - Escherichia coli (strain K-12) | 32317 | 127 | 6.3E-09 | 9 | 41% |
| | Q1R6A3_ECOUT | Malate dehydrogenase (EC 1.1.1.37).- Escherichia coli (strain UTI89 / UPEC). | 35036 | 125 | 1E-08 | 9 | 38% |
| | Q9ETZ1_ECOLI | Malate dehydrogenase (Fragment).- Escherichia coli. | 30099 | 107 | 6.30E-07 | 8 | 40% |
| | Q9ETZ7_ECOLI | Malate dehydrogenase (Fragment).- Escherichia coli. | 30086 | 107 | 6.3e-07 | 8 | 40% |
| | Q9F6J4_ECOLI | Malate dehydrogenase (Fragment).- Escherichia coli. | 30102 | 107 | 6.30E-07 | 8 | 40% |
| P0A9B2 | Q1RB13_ECOUT | Glyceraldehyde-3-phosphate dehydrogenase A (EC 1.2.1.12).- Escherichia coli (strain UTI89 / UPEC). | 35933 | 98 | 5.70E-06 | 9 | 28% |
| | DEECG3 | glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (EC 1.2.1.12) A - Escherichia coli (strain K-12) | 35510 | 98 | 5.70E-06 | 9 | 28% |
| | G3P1_ECO57 | Glyceraldehyde-3-phosphate dehydrogenase A (EC 1.2.1.12) (GAPDH-A).- Escherichia coli O157:H7. | 35379 | 97 | 6.8E-06 | 9 | 28% |
| | AAC43271 | ECU07750 NID: - Escherichia coli | 33507 | 80 | 0.0003 | 8 | 27% |
| | 1GAEO | D-glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) mutant N313T holo form, chain O - Escherichia coli | 35366 | 78 | 0.00047 | 8 | 24% |
| | AAA23838 | ECOGAPAB NID: - Escherichia coli | 33012 | 64 | 0.014 | 7 | 23% |
| P0AFL3 | CAB69331 | SEQUENCE 1 FROM PATENT WO9845454 (fragment).- unidentified. | 18066 | 112 | 2E-07 | 6 | 64% |
| | AAA24261 | ECOPABAA NID: - Escherichia coli | 16847 | 87 | 5.80E-05 | 5 | 55% |
| | CSECA | peptidylprolyl isomerase (EC 5.2.1.8) A precursor - Escherichia coli (strain K-12) | 20418 | 83 | 0.00017 | 5 | 44% |
| P0A6P9 | AAA24486 | ECOPYRG NID: - Escherichia coli | 12523 | 49 | 4.10E-01 | 3 | 49% |
| | CAA57795 | ECENO NID: - Escherichia coli | 46417 | 32 | 18 | 3 | 13% |
| | ENO_ECO57 | Enolase (EC 4.2.1.11) (2-phosphoglycerate dehydratase) (2-phospho-D-glycerate hydro-lyase).- Escherichia coli O157:H7. | 45495 | 32 | 18 | 3 | 14% |
| P00811 | 2BLSA | ampc beta-lactamase (EC 3.5.2.6), chain A - Escherichia coli | 39398 | 101 | 2.5e-06 | 8 | 28% |
| | 1C3BA | cephalosporinase (EC 3.5.2.6), chain A - bacteria | 39526 | 101 | 2.50E-06 | 8 | 28% |
| | Q6PRU8_ECOLI | Beta-lactamase (Fragment). - Escherichia coli. | 38673 | 80 | 0.0003 | 7 | 25% |
| | Q5YEX8_ECOLI | Extended-spectrum beta lactamase. - Escherichia coli. | 41573 | 79 | 0.00039 | 7 | 24% |
| | Q5YEX7_ECOLI | Extended-spectrum beta lactamase. - Escherichia coli. | 41654 | 79 | 0.00039 | 7 | 24% |

**Note:** Gray color is the correct theoretical protein of each AC number. Black boxes represent the results of each AC number that the correct theoretical protein was found within the third order.

**Table 3b. 'Hits' Results from the Current *Spirulina* Tool Using PMF Data of *E. coli K-12 substr. DH10B***

| AC No. | # | Protein ID | Orf Name | Computer Annotation | Human Curation | Score | Match | pI (Wiki) | Protein MW | % Coverage | Fragments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P0C0V0 | 1 | 135 | YP_001729118.1 | serine endoprotease (protease Do), membrane-associated [Escherichia coli str. K-12 substr. DH10B] | | 10.7 | 6/9 | 8.7888 | 49305.4498 | 19.6203 | 6 |
| | 2 | 2368 | YP_001731351.1 | sulfate/thiosulfate ABC transporter ATP-binding protein [Escherichia coli str. K-12 substr. DH10B] | | 11.05 | 4/9 | 7.1262 | 41015.4896 | 16.7123 | 4 |
| | 3 | 3468 | YP_001732451.1 | lipopolysaccharide core biosynthesis [Escherichia coli str. K-12 substr. DH10B] | | 11.2 | 2/9 | 8.6790 | 41684.9217 | 13.7255 | 4 |
| | 4 | 835 | YP_001729818.1 | D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 6a) [Escherichia coli str. K-12 substr. DH10B] | | 11.24 | 2/9 | 8.1077 | 43563.34 | 11 | 3 |
| | 5 | 2137 | YP_001731120.1 | oligopeptide ABC transporter membrane protein [Escherichia coli str. K-12 substr. DH10B] | | 11.25 | 2/9 | 8.5691 | 40316.5402 | 8.2418 | 2 |
| P61889 | 1 | 3105 | YP_001732088.1 | malate dehydrogenase, NAD(P)-binding [Escherichia coli str. K-12 substr. DH10B] | | 9.91 | 8/9 | 5.37 | 32299.205 | 41.0256 | 8 |
| | 2 | 2814 | YP_001731797.1 | methylmalonyl-CoA decarboxylase, biotin-independent [Escherichia coli str. K-12 substr. DH10B] | | 10.66 | 4/9 | 5.67 | 29135.9555 | 24.5211 | 4 |
| | 3 | 1689 | YP_001730672.1 | quinate/shikimate 5-dehydrogenase, NAD(P)-binding [Escherichia coli str. K-12 substr. DH10B] | | 10.99 | 3/9 | 4.76 | 31189.7272 | 12.1528 | 3 |
| | 4 | 175 | YP_001729158.1 | 2,5-diketo-D-gluconate reductase B [Escherichia coli str. K-12 substr. DH10B] | | 11.05 | 3/9 | 5.28 | 29400.4476 | 11.6105 | 3 |
| | 5 | 2468 | YP_001731451.1 | 3-mercaptopyruvate sulfurtransferase [Escherichia coli str. K-12 substr. DH10B] | | 11.08 | 2/9 | 4.29 | 30774.5998 | 8.1851 | 2 |
| P0A9B2 | 1 | 1772 | YP_001730755.1 | glyceraldehyde-3-phosphate dehydrogenase A [Escherichia coli str. K-12 substr. DH10B] | | 13.28 | 6/11 | 6.71 | 35492.2316 | 21.7523 | 6 |
| | 2 | 2881 | YP_001731864.1 | hydrogenase 2 4Fe-4S ferredoxin-type component [Escherichia coli str. K-12 substr. DH10B] | | 13.54 | 4/11 | 7.00 | 35961.5596 | 13.4146 | 4 |
| | 3 | 1881 | YP_001730864.1 | hypothetical protein ECDH10B_2030 [Escherichia coli str. K-12 substr. DH10B] | | 13.61 | 4/11 | 9.52 | 34146.682 | 12.987 | 4 |
| | 4 | 2225 | YP_001731208.1 | peptidase [Escherichia coli str. K-12 substr. DH10B] | | 13.63 | 2/11 | 5.17 | 35890.1515 | 7.7399 | 3 |
| | 5 | 4023 | YP_001733006.1 | aspartate carbamoyltransferase, catalytic subunit [Escherichia coli str. K-12 substr. DH10B] | | 13.65 | 2/11 | 6.15 | 34387.7513 | 11.8971 | 4 |

**(Table 3b). contd…..**

| AC No. | # | Protein ID | Orf Name | Computer Annotation | Human Curation | Score | Match | pI (Wiki) | Protein MW | % Coverage | Fragments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| P0AFL3 | 1 | 3218 | YP_001732201.1 | peptidyl-prolyl cis-trans isomerase A (rotamase A) [Escherichia coli str. K-12 substr. DH10B] | | 6.135 | 4/6 | 9.35 | 20400.3836 | 42.6316 | 4 |
| | 2 | 633 | YP_001729616.1 | apo-citrate lyase phosphoribosyl-dephospho-CoA transferase [Escherichia coli str. K-12 substr. DH10B] | | 7.247 | 2/6 | 6.86 | 20239.6361 | 27.3224 | 2 |
| | 3 | 532 | YP_001729515.1 | apo-citrate lyase phosphoribosyl-dephospho-CoA transferase [Escherichia coli str. K-12 substr. DH10B] | | 7.247 | 2/6 | 6.86 | 20239.6361 | 27.3224 | 2 |
| | 4 | 2615 | YP_001731598.1 | glucitol/sorbitol-specific enzyme IIC component of PTS [Escherichia coli str. K-12 substr. DH10B] | | 7.394 | 1/6 | 8.33 | 20548.7225 | 11.7647 | 1 |
| | 5 | 958 | YP_001729941.1 | methylglyoxal synthase [Escherichia coli str. K-12 substr. DH10B] | | 7.394 | 1/6 | 6.19 | 16889.7409 | 13.8158 | 1 |
| P0A6P9 | 1 | 2542 | YP_001731525.1 | 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase, tyrosine-repressible [Escherichia coli str. K-12 substr. DH10B] | | 5.859 | 2/5 | 5.25 | 38761.4192 | 16.573 | 4 |
| | 2 | 2688 | YP_001731671.1 | enolase [Escherichia coli str. K-12 substr. DH10B] | | 6.078 | 3/5 | 5.08 | 45608.4117 | 14.1204 | 3 |
| | 3 | 3530 | YP_001732513.1 | hypothetical protein ECDH10B_3875 [Escherichia coli str. K-12 substr. DH10B] | | 6.15 | 1/5 | 5.82 | 44938.227 | 6.9307 | 2 |
| | 4 | 3774 | YP_001732757.1 | UDP-N-acetylenolpyruvoylglucosamine reductase, FAD-binding [Escherichia coli str. K-12 substr. DH10B] | | 6.15 | 1/5 | 5.74 | 37809.2744 | 7.6023 | 2 |
| | 5 | 3609 | YP_001732592.1 | entero common antigen (ECA) polysaccharide chain length modulator [Escherichia coli str. K-12 substr. DH10B] | | 6.153 | 2/5 | 6.23 | 39445.9437 | 12.3563 | 2 |
| P00811 | 1 | 3937 | YP_001732920.1 | beta-lactamase/D-alanine carboxypeptidase [Escherichia coli str. K-12 substr. DH10B] | | 8.845 | 6/8 | 9.08 | 41511.4022 | 27.0557 | 8 |
| | 2 | 3468 | YP_001732451.1 | lipopolysaccharide core biosynthesis [Escherichia coli str. K-12 substr. DH10B] | | 9.949 | 3/8 | 8.68 | 41684.9217 | 11.4846 | 3 |
| | 3 | 789 | YP_001729772.1 | ABC transporter membrane protein [Escherichia coli str. K-12 substr. DH10B] | | 9.971 | 2/8 | 8.16 | 42014.5599 | 6.1008 | 2 |
| | 4 | 2968 | YP_001731951.1 | sodium:serine/threonine symporter [Escherichia coli str. K-12 substr. DH10B] | | 9.973 | 1/8 | 8.27 | 43431.3994 | 7.4879 | 2 |
| | 5 | 1999 | YP_001730982.1 | lipopolysaccharide biosynthesis protein [Escherichia coli str. K-12 substr. DH10B] | | 10.03 | 2/8 | 9.32 | 43142.6376 | 8.3333 | 3 |

loaded to test our in-house tool and also to compare with Mascot by using the PMF data of *E. coli* proteins (cross-species proteins), P0C0V0-protease Do, P61889-malatede hydrogenase, P0A9B2-glyceraldehyde-3-phosphate dehydrogenase A, P0AFL3-peptidyl-prolyl cis-trans isomerase A, P0A6P9- enolase and P00811-beta-lactamase as the input PMFs. The best 'Hits' results from Mascot and our current tool are shown in Table **3a** and Table **3b**, respectively.

According to the search results from Mascot, these proteins were found first in the best 'Hits' results for the first three proteins, and third for the last three, as shown in gray boxes in Table **3a**. The input parameters, MSDB (the Mascot database), taxonomy of *E. coli*, an allowed miss cleavage of one, and a peptide tolerance of 1.2 Da were used.

The results from our current tool illustrated that five out of six proteins were shown first in the 'Hits' results. The protein with accession number P0A6P9 was identified second on the list as enolase (protein ID 2688) with a score of 6.0782, which is less than that of the 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase.

In conclusion, the current tool has an accurate ability to identify proteins using the scoring function and appropriate input parameters. This version of the tool with a probability-based scoring function has high accuracy in protein identification by using peptide mass fingerprints. To the best of our knowledge, this is the first time that a tool used as search engine for protein identification contains pI-filtering, pI-calculating and a batch processing module. When the limitation of batch processing is its long execution time, this problem can be overcome by automatically obtaining the protein mass and pI values from each PMF file of the batch process. Improvements to the batch process are in progress. Thus, the current tool obtained in this study has a high impact on the protein identification step in our proteomic work due to its accurate scoring function and ranking criteria, including pI.

## ACKNOWLEDGEMENTS

This research was funded by a grant from the King Mongkut's University and Technology of Thonburi and the National Center for Genetic Engineering and Biotechnology (BIOTEC), Bangkok, Thailand.

## SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

## REFERENCES

[1] J. C. Boisson, L. Jourdan, E. G. Talbi, and C. Rolando, "Protein sequencing with an adaptive genetic algorithm from tandem mass spectrometry," *in 2006 IEEE Congress on Evolutionary Computation, CEC* 2006, pp. 1412-1419.

[2] Y. Wan and T. Chen, "A hidden Markov model based scoring function for mass spectrometry database search," *in Lecture Notes in Bioinformatics (Subseries of Lecture Notes in Computer Science)*, 2005, pp. 342-356.

[3] J. K. Eng, A. L. McCormack, and J. R. Yates Iii, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 976-989, 1994.

[4] A. Sillero and A. Maldonado, "Isoelectric point determination of proteins and other macromolecules: Oscillating method," *Computers in Biology and Medicine*, vol. 36, pp. 157-166, 2006.

[5] K. Gevaert and J. Vandekerckhove, "Protein identification methods in proteomics," *Electrophoresis*, vol. 21, pp. 1145-1154, 2000.

Received: June 25, 2009 Revised: August 14, 2009 Accepted: August 21, 2009