Genomics & Informatics Vol. 4(2) 87-93, June 2006

Improved Algorithms for the Identification of Yeast Proteins and Significant Transcription Factor and Motif Analysis

Seung-Won Lee[1], Seong-Eui Hong[2], Kyoo-Yeol Lee[1], Do-il Choi[1], Hae-Young Chung[2] and Cheol-Goo Hur[1]*

[1]Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-333, Korea, [2]Pusan National University, Pusan 609-735, Korea

Abstract

With the rapid development of MS technologiesy, the demands for a more sophisticated MS interpretation algorithm haves grown as well. We have developed a new protein fingerprinting method using a binomial distribution, (fBIND). Withthe fBIND, we improved the performance accuracy of protein fingerprinting up to the maximum 49% (more than MOWSE) and 2% than(at a previous binomial distribution approach studied by of Wool *et al.*) as compared to the established algorithms. Moreover, we also suggest a the statistical approach to define the significance of transcription factors and motifs in the identified proteins based on the Gene Ontology (GO).

Abbreviations: fBIND, fingerprinting using binomial distribution; GO, Gene Ontology; MS, Mass Spectrometry; PMF, peptide mass fingerprinting; nr, nonredundant; SGD, Saccharomyces Genome Database

Keywords: peptide mass fingerprinting, molecular weight search, binomial distribution, hypergeometric distribution

Introduction

As the tendency in biology researches has been moved from analysis on few genes/proteins to macroanalysis on more extensive genes/proteins, MS has been recognized as key biotechnology. In particular, MS can be considered as the most important tool in performing proteomics to understand biological phenomenon of living organisms at the large-scale protein level. Accordingly, a number of recent researches presented several software and algorithms that can effectively analyze the results of MS and relevant researches have been actively continued (Fenyo, 2000; rogers *et al.*, 2003; Cutler *et al.*, 2003).

While PepSea and PeptIdent/MultiIdent identified proteins on the basis of the number of matches between peptide molecular weight in database and peptide from experiments, they didn't support accurate analysis (Mann *et al.*, 1993; Wilkins *et al.*, 1999; Wilkins *et al.*, 1998). MOWSE used scoring method that reflected characteristics of database beyond simple matches as calculating frequencies of peptide generated from proteins at 10kD intervals (Rappin *et al.*, 1993). The researches based on probability are MASCOT that identified more reliable analysis as calculating the probability that random hit could occur, ProFound that calculated probability values of proteins using Bayesian approach and research by Wool *et al.* that made scoring method using binomial distribution (Perkins *et al.*, 1999; Zhang *et al.*, 2000; Wool and Smilansky, 2002; Clauser *et al.*, 1999). MassSorter which provides visual and user-friendly interface is a convenient tool to analyze data from MS experiment (Barsners *et al.*, 2006).

While MOWSE scoring algorithm that is widely used in general is simple and relatively accurate, it has the limit to depend on relatively large masses among peptide fragments in experiments. In accordance with the research by Pappin *et al.*, most proteins are distributed around 50kD and peptides created as hydrolyzing these proteins in trypsin are mainly small fragments. As reported above, yeast proteins also showed the similar distribution (data not shown). While most peptides from experiments are small fragments, fragments don't have substantial effects on MOWSE score, but a small number of fragments with relatively large molecular weight fragment significantly influence on MOWSE score. As shown in Fig. 1(A), for the performance of MOWSE scoring, its accuracy is substantially decreased when molecular weights of peptides are small. This phenomenon is also appeared according to intact molecular weight of protein as shown in Fig. 1(B). The reason is that proteins with larger molecular weights create more peptide fragments and on this occasion, a number of fragments with lower molecular weights are generated. This MOWSE scoring algorithm is also applied to other peptide mass fingerprinting (PMF) programs such as MASCOT and MS-Fit. Although scoring algorithm in the research by Wool *et al.* was based on relatively simple probability, its performance was very remarkable as

*Corresponding author: E-mail hurlee@kribb.re.kr, Tel +82-42-879-8560, Fax +82-42-879-8569

Accepted by 31 May 2006

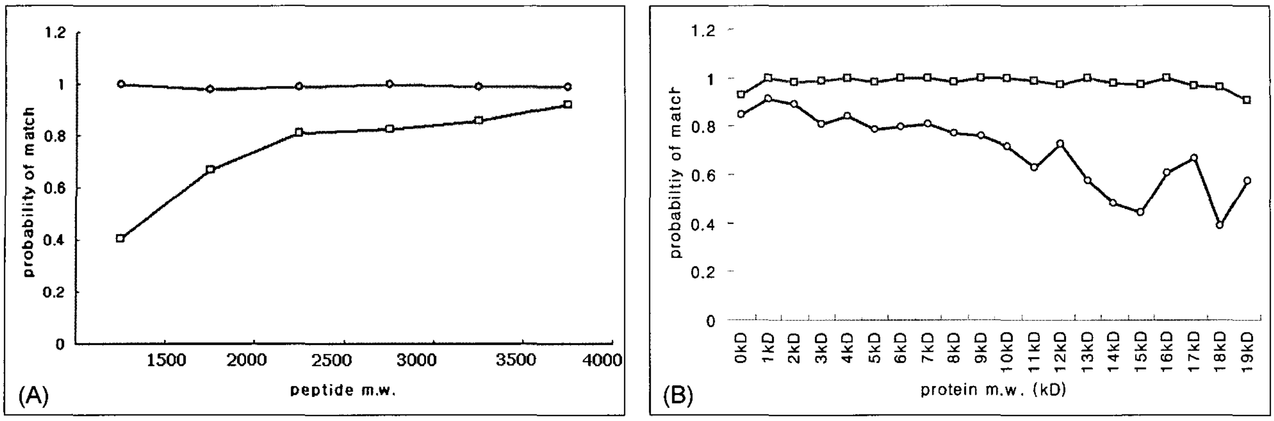

Fig. 1. Performance Comparison of MOWSE and Wool *et al.*'s study. (A) Performance comparison of MOWSE (squares) and Wool *et al.*'s study (circles) according to peptide ranges using SWISS- PROT release 42; (B) Performance comparison of MOWSE (squares) and Wool *et al.*'s study (circles) according to protein molecular weight using SWISS-PROT release 42.

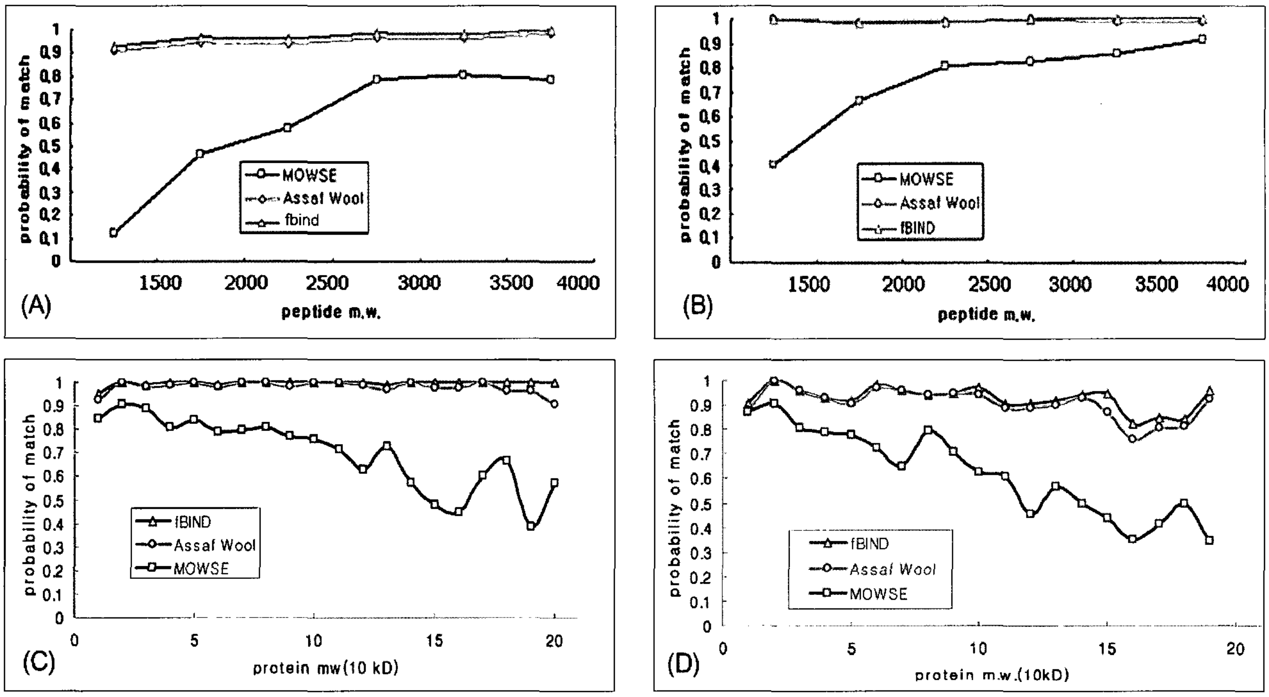

Fig. 2. Performance Comparison of Algorithms. (A) Performance of MOWSE (squares) and fBIND (triangles) in accordance with peptide ranges using SWISS-PROT release 42; (B) Performance of MOWSE (squares), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance with peptide ranges using nr; (C) Performance of MOWSE (squares), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance with protein molecular weight using SWISS-PROT release 42; (D) Performance of MOWSE (squares), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance protein molecular weight using nr.

shown in Fig. 1. It demonstrated uniform accuracies regardless of peptide fragment ranges and intact molecular weights of proteins as shown in Fig. 1. However, its accuracy tended to be slightly declined because of repetitive sequences. Accordingly, this study developed fBIND to identify proteins more accurately as making up for disadvantages of existing algorithms.

Methods and Results

Peptide Search Database

Peptide pool to analyze PMF was generated using SWISS-PROT release 42 and nonredundant (nr) of NCBI. Then, only yeast proteins were collected in SWISS-PROT 42 and NCBI nr and then, theoretical

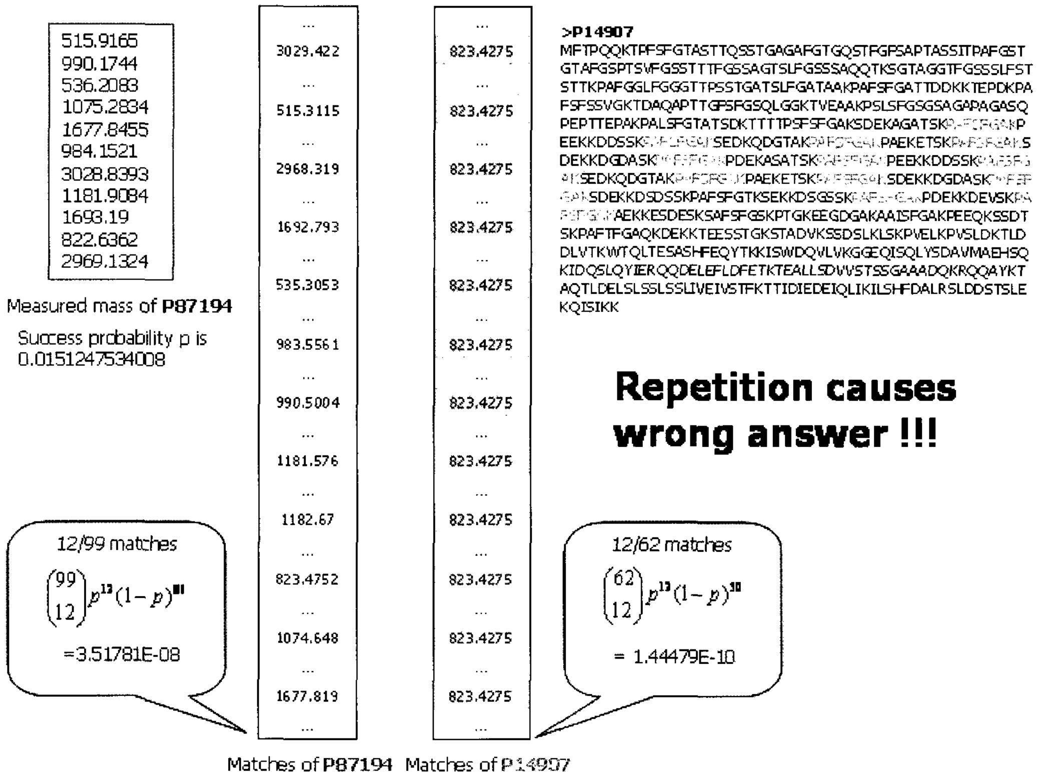
515.9165
990.1744
536.2083
1075.2834
1677.8455
984.1521
3028.8393
1181.9084
1693.19
822.6362
2969.1324
Measured mass of P87194
Success probability p is 0.0151247534008
12/99 matches
=3.51781E-08
...
3029.422
515.3115
2968.319
1692.793
535.3053
983.5561
990.5004
1181.576
1182.67
823.4752
1074.648
1677.819
823.4275
>P14907
MFTPQQKTPFSFGTASTTQSSTGAGAFGTGQSTFGPSAPTASSITPAFGST
GTAFGSPTSVFGSSTTTFGSSAGTSLFGSSSAQQTKSGTAGGTFGSSSLFST
STTKPAFGGLFGGGTTPSSTGATSLFGATAAKPAFSFGATTDDKKTEPDKPA
FSFSSVGKTDAQAPTTGPSFGSQLGGKTVEAAKPSLSFGSGSAGAPAGASQ
PEPTTEPAKPALSFGTATSDKTTTTPSFSFGAKSDEKAGATSK
EEKKDDSSK SEDKQDGTAK PAEKETSK S
DEKKDGDASK PDEKASATSK PEEKKDDSSK
SEDKQDGTAK PAEKETSK SDEKKDGDASK
SDEKKDSDSSKPAFSFGTKSEKKDSGSSK PDEKKDEVSK
AEKKESDESKSAFSFGSKPTGKEEGDGAKAAISFGAKPEEQKSSDT
SKPAFTFGAQKDEKKTEESSTGKSTADVKSSDSLKLSKPVELKPVSLDKTLD
DLVTKWTQLTESASHFEQYTKKISWDQVLVKGGEQISQLYSDAVMAEHSQ
KIDQSLQYIERQQDELEFLDFETKTEALLSDVVSTSSGAAADQKRQQAYKT
AQTLDELSLSSLSSLIVEIVSTFKTTIDIEDEIQLIKILSHFDALRSLDDSTSLE
KQISIKK
Repetition causes wrong answer !!!
12/62 matches
= 1.44479E-10
Matches of P87194
Matches of P14907

Fig. 3. Repetitive peptides in a protein. Protein p14907 was searched by peptides extracted from protein p87194 with the binomial distribution approach. This incorrect result arose from many Repetitive peptides.p14907, p 87194

peptide pool created by trypsin was built up. This peptide pool consisted of only peptides in 500-4000Da that was experimentally significant among molecular weights of peptides. Moreover, in consideration of errors caused by biological experiments, this study assumed the missed cleavage level of trypsin to level 2 and peptides generated on this level were included in peptide pool. Next, peptide pool generated by the process above and data related to yeast protein were used to construct database using MySQL 4.2.13. Yeast proteins in the database were 5,406 for SWISS-PROT 42 and 9,007 for NCBI nr. Peptide pool generated by trypsin consisted of 718,106 and 1,040,173 records for SWISS-PROT 42 and NCBI nr, respectively.

Improved algorithm for identifying a protein

Like the research by Wool *et al.* scoring method in fBIND calculated probabilities with respect to binomial distribution as described below and applied overlap penalty to reduce impacts from repetitive sequences (Fig. 3) and proportional relation between molecular weights of proteins and the number of peptides.

$$P(N,r)=\binom{N}{r}p^{r}(1-p)^{N-r}\times e^{k}$$

N = total number of peptides in a protein
r = number of random match
p = number of match / number of peptide in database
k = frequency of overlapping match in a protein

Performance of fBIND

The results by fBIND were compared to those by MOWSE and researches by Wool *et al.* in order to evaluate performances of fBIND that were modified above. The test sets for analysis and comparison of performance randomly selected 100 proteins from SWISS-PROT release 42 and then 10 peptides from each theoretical peptide pool. Next, random error values among e = {-0.999, -0.998, ⋯, 0, ⋯, 0.998, 0.999} in the calculated mass of each selected peptide were added up for the test sets. These test sets proteins were divided into the sets by molecular weight of eacha peptide (<1500Da,

Table 1. Description about function specificity of motifs of protein that is within a particular function category and comparison to InterPro GO mapping

Accession number	Function category	Motif	Function category-specificity	Motif-related other function	InterPro GO
O13527	DNA recombination	Integrase, catalytic domain	YES	NONE	DNA recombination
O13528	DNA recombination	TYA transposon protein	YES	NONE	NONE
P00330	Energy pathways	Zinc-containing alcohol dehydrogenase superfamily	YES	alcohol metabolism aldehyde metabolism	NONE
		Zinc-containing alcohol dehydrogenase	YES	alcohol metabolism aldehyde metabolism	NONE
P00410	Energy pathways	Copper center Cu(A)	NO	NONE	NONE
		Cytochrome c oxidase, subunit II	NO	NONE	electron transport
		Cupredoxin	YES	ion transport response to abiotic stimulus	NONE

1500~ 2000Da, 2000~2500Da, 2500~3000Da, 3000~3500Da, 3500~4000Da) and that by molecular weight (<10kD, 10~20kD, ···, 180~190kD, 190~200kD) of eacha protein, and each set was independently analyzed. It was assumed that missed cleavage was 0, mass tolerance was 1Da and there was no modification.

Consequently, as shown in Fig. 2(a, b), MOWSE showed low accuracies of 40.4% (SWISS-PROT release 42) and 12% (nr) in PMF with peptides of less than 1500Da. The accuracies of MOWSE were gradually improved in accordance with increases of molecular weights of peptides. For peptides below 4000 Da, the performances of SWISS-PROT release 42 and nr were improved up to 92% and 79%, respectively. This phenomenon in accordance with peptide ranges also influenced on performances according to intact protein molecular weights. As shown in Fig. 2(c, d), as the sizes of proteins became larger, performances were gradually decreased.

Then, in the range between 190kD and 200kD, the accuracies in SWISS-PROT release 42 and nr reached 57% and 31%, respectively. As shown in Fig. 2, the performances of fBIND didn't show significant differences as compared to researches by Wool *et al.* It is because scoring method of Wool *et al.* showed high performances of 99.1% and 97.9% in average in accordance with peptide ranges and protein ranges, respectively, when SWISS-PROT release 42 was applied, as shown in Fig. 1. The fBIND that considered influences by increases of random hit rates in accordance with increases of peptide fragments and repetitive sequences demonstrated higher performances as compared to researches by Wool *et al.* As illustrated in Fig. 3 (a, b), the performances in accordance with peptide ranges were 99.6% and 97.1% for SWISS-PROT release 42 and nr, respectively, with increases of about 1% and 2% for SWISS-PROT release 42 and nr, respectively, as compared to study of Wool *et al.* The performances in accordance with sizes of proteins were also improved. The performances were 99.6% and 90.9% in average for SWISS-PROT release 42 and nr, respectively, with increases of up to 10% and 7% for SWISS-PROT release 42 and nr, respectively, as compared to researches by Wool *et al.*

Table 2. The number of mapping proteins

Method	Mapping count
INTERPRO	2211
Cumulative hypergeometric probability	2223
Total count of annotated proteins	3120

Significant transcription factor/motif analysis

This study tried to identify proteins more accurately as improving existing algorithms and analyzed significant transcription factor/motif analysis to provide useful information on identified proteins. MATCH^{TM} and PATCH ^{TM} of TRANSFAC, representative transcription factor analysis tools, are very useful to find out transcription factors binding on specific sequences using position-specific matrixes and patterns (Matys *et al.*, 2003; Kel *et al.*, 2003). Moreover, InterPro, the representative motif database, is important database collecting motifs in each protein (Mulder *et al.*, 2003). However, all of them don't provide information on how much important it is the contribution of transcription factors or motifs identified by those tools and database on functions. In accordance with analysis of transcription factors existing on upstream of yeast ORF by MATCH^{TM} and PATCH^{TM}, it was observed that a number of transcription factors were abundantly appeared regardless of specific functions including that HIF-1 was appeared up to 1614 times on total 5,406 upstream. Therefore, this study tried to analyze significant transcription factors and motifs contributing on specific functions using cumulative hypergeometric probability distribution as described below.

Process of transcription factors and motifs analysis

First of all, yeast ORF and proteins were divided into relatively detailed function categories of about 105 with reference to data annotated with respect to the process among GO terminology in Saccharomyces Genome Database (SGD) (Dwight *et al.*, 2002). Then, mapping of ORF region on chromosome was conducted using sim4 to acquire transcription factors binding on upstream of ORF included in each function category (Florea *et al.*, 1998). As reported by Zhu *et al.*, -1000 regions from a translation start site was considered as upstream region and transcription factors in relevant upstream region were analyzed by MATCH^{TM} and PATCH^{TM} (Zhu and Zhang, 1999). Motifs of each protein were acquired from InterPro database. Transcription factors and frequencies of motifs in each function category segmented into 105 categories were calculated and then, transcription factors and motifs characteristically appeared in each function category were analyzed through cumulative hypergeometric probability distribution as shown below.

$$P\{x=i\}=\frac{\binom{m}{i}\binom{N-m}{r-i}}{\binom{N}{r}}$$

N = Total number of ORF/proteins

r = Number of ORF/proteins in a specific category

m = Number of specific transcription factors/motifs identified in total ORF/proteins

i = Number of specific transcription factors/motifs identified in ORF/proteins in a specific category

$$P\{X\geq i\}=\sum_{j=i}^{r}P\{X=j\}$$

When $p\{X\geq i\}\leq 0.001$ satisfied, relevant motif is considered as the motifs specifically generated in specific function categories. Table 1 described the examples of transcription factors and motifs specifically generated in each function acquired by the process above.

Estimation for analyzing transcription factors and motifs

Table 1 described motifs and relevant characteristics existing in each protein. O13527 has integrase motif as the protein included in DNA recombination among 105 segmented categories. As a result of analysis using cumulative hypergeometric probability distribution, integrase catalytic domain is a specific motif that is appeared especially a lot in DNA recombination category and it is considered that it contributes on the functions related to DNA recombination of O13527. This result is the same as GO mapping results of InterPro. P00330 falls under GO:0006113 fermentation according to GO mapping of SGD and includes in energy pathway, the upper category. Zinc-containing alcohol dehydrogenase superfamily, a motif found in P00330, is especially appeared a lot in energy pathway category, alcohol metabolism and aldehyde metabolism category. It contributes on functions related to energy pathway of P00330 and has the possibility to take part in other functions such as alcohol metabolism and aldehyde pathway. InterPro doesn't provide information on this motif. In accordance with GO mapping of SGD, P00410 was mapped to GO:0009060 aerobic respiration and included in energy pathway, the upper category. Copper center cu(A), an identified motif, was a specific motif neither to energy pathway category nor to other categories. However, cupredoxin, another motif, was appeared especially in the same category as P00410 and also the specific motif in ion transport and response to abiotic stimulus category. As explained above, this study provided information about contribution of motifs identified in each protein through cumulative hypergeometric probability distribution on functions and suggested to broaden function annotation more extensively. In accordance with annotation of yeast protein with respect to process among GO mapping data of InterPro, about 53% of total yeast proteins could be annotated. Meanwhile, when significant factor/motif analysis of this study was combined with InterPro, about 83% of total proteins were covered (Table 2). The detailed GO mapping information is available at http://genepool.kribb.re.kr/pmf/ coverage.txt

Discussion

As large-scale protein researches have been analyzed, researches on instrument related to MS, experimental techniques and PMF algorithm have been actively studied. In particular, we need the algorithm to identify accurate proteins without being sensitive to experimental errors in order to identify effective proteins. In accordance with the development of MS, a wide range of research results from algorithms of simple matches to algorithms based on sophisticated probabilities has been reported. Each algorithm demonstrates unique characteristics in accordance with scoring methods. For example, MOWSE scoring method based on frequency that consists of mainly peptide fragments show lower performances, but that with peptide fragments of high molecular weights demonstrates significantly higher performances.

Moreover, for PMF based on binomial distribution presented in the researches by Wool *et al.* the performance tends to be slightly decreased when measured peptide is randomly matched to proteins with repetitive sequences.

Consequently, this study developed the system showing better performances as making up for scoring methods based on MOWSE and binomial distribution among PMF algorithms that have been studied until now. Furthermore, this study tried to provide information for interpreting peptide fingerprinting results for researchers as analyzing significant factors/ motifs related to regulations and activities of identified proteins as well as for accurate identification of proteins. In the future, this study will provide information on fBIND with respect to protein-protein interaction data and protein sub-cellular localization information. The fBIND is available at http:// genepool.kribb.re.kr/pmf/ and supplementary information is available at http://genepool.kribb.re.kr/pmf/Instruction/ instruction.html

Acknowledgements

This research was supported by a grant (CGS0110424) from Crop Functional Genomics Center.

References

Barsnes, H., Mikalsen, S.O., and Eidhammer, I. (2006). MassSorter: a tool for administrating and analyzing data from mass spectrometry experiments on proteins with known amino acid sequences. *BMC bioinformatics* 7, 42.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* 13, 662-672.

Clauser, K.R., Baker, P., and Burlingame, A.L. (1999). Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71, 2871-2882.

Cutler, P., Heald, G., White, I.R., and Ruan, J. (2003). A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection. *Proteomics* 3, 392-401.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J.M. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69-72.

Fenyo, D. (2000). Identifying the proteome. *Curr. Opin. Biotechnol.* 11, 391-395.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence, *Genome Res.* 8, 967-974.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E, Kel-Margoulis O.V., and Wingender E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31, 3576-3579.

Mann, M., Hojrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338-344.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374-378.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R., and Zdobnov, E.M. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* 31, 315-318.

Pappin, D.J., Hojrup, P., and Bleasby, A.J. (1993). Rapid identification of proteins by peptide-mass finger printing. *Curr. Biol.* 3, 327-332.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551-3567.

Rogers, M., Graham, J., and Tonge, R.P. (2003). Statistical Moddels of Shape for the Analysis of Protein Spots in 2-D Electrophoresis Gel Images. *Proteomics* 3, 879-886.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Willianms, K.L., Appel, R.D., and Hochstrasser, D.F. (1999). Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* 112, 531-552.

Wilkins, M.R., Gasteiger, E., Wheeler, C.H., Lindskog, I., Sanchez, J., Bairoch, A., Appel, R.D., Dunn, M.J., and Hochstrasser D.F. (1998). Multiple parameter cross-species protein identification using MultiIdent-a world-wide web accessible tool. *Electrophoresis* 19, 3199-3206.

Wool, A. and Smilansky, Z. (2002). Precalibration of matrix-assisted laser desorption/ ionization-time of flight spectra for peptide mass fingerprinting. *Proteomics* 2, 1365-1373.

Zhang, W. and Chait, B.T. (2000). ProFound- an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* 72, 2482-2489.

Zhu, J. and Zhang, M.Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 15, 607-611.