Biochemical and Biophysical Research Communications 377 (2008) 253–256

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

ASPMF: A new approach for identifying alternative splicing isoforms using peptide mass fingerprinting ☆

Seung-Won Lee, Jae-Pil Choi, Hyun-Jin Kim, Ji-Man Hong, Cheol-Goo Hur *

Omics and Integration Research Center, KRIBB, 52Eoeun-dong, Yuseong-gu, Daejeon 305-333, Republic of Korea

ARTICLE INFO

Article history:
Received 17 September 2008
Available online 1 October 2008

Keywords:
Alternative splicing
Peptide mass fingerprinting (PMF)
Protein isoform
Protein identification
Mass spectrometry (MS)
Tandem MS
MS/MS

ABSTRACT

Alternative splicing is generally accepted as a mechanism that explains the discrepancy between the number of genes and proteins. We used peptide mass fingerprinting with a theoretical database and scoring method to discover and identify alternative splicing isoforms. Our theoretical database was built using published alternative splicing databases such as ECgene, H-DBAS, and TISA. According to our theoretical database of 190,529 isoforms, 37% of human genes have multiple isoforms. The isoforms produced from a gene partially share common peptide fragments because they have common exons, making it difficult to distinguish isoforms. Therefore, we developed a new method that effectively distinguishes a true isoform among multiple isoforms in a gene. In order to evaluate our algorithm, we made test sets for 4226 protein isoforms extracted from our theoretical database randomly. Consequently, 94% of true isoforms were identified by our scoring algorithm.

About 40–60% of human genes, and more than 20% of Arabidopsis genes, undergo alternative splicing. Alternative splicing is now largely accepted as one of the mechanisms that accounts for the discrepancy between the high-level complexity of humans and their relatively small number of genes [1]. Alternative splicing also influences several diseases and gene function [2–7].

As alternative splicing mechanisms have become available, alternative splicing variants are being predicted with statistical techniques and biological data. Collectively, these variants have been used to build public alternative splicing databases. Alternative splicing databases, including TISA [8], ECgene [9], and H-DBAS [10], were developed by graphical methods that visually indicate possible exons in conjunction with and biological data, such as Expressed Sequence Tags (ESTs) and full-length cDNA. These predictive transcript databases contain various numbers of isoforms; for example, there are 97,286 isoforms in TISA, 30,389 in H-DBAS and 185,174 in ECgene. All of these three databases are associated with human splicing, but there are also alternative spicing databases for Arabidopsis and rice based on EST and cDNA data [11,12]. However, the putative transcripts should be experimentally validated because many transcripts within databases are based only on bioinformatics. Many researchers have been working to validate spliced transcripts and exons experimentally. Tanner et al. utilized tag sequences, short protein sequences identified by mass spectrometry (MS) to determine splicing events and isoforms after a putative exon database was constructed with tools that predict exon [13]. A genome-wide approach mapped the masses from MS experiments on the *Escherichia coli* chromosome to find frameshifts, incorrect protein annotations, and alternative splicing events [14]. Tissue- and disease-specific transcripts and alternative splicing patterns from a gene can be detected by exon, tiling, or exon junction microarrays [15–17]. Brain- and heart-specific alternative splicing transcripts were identified by microarray experiments and RT-PCR [16].

Peptide mass fingerprinting (PMF) has been widely utilized to identify proteins. Protein samples of interest are usually isolated by two-dimensional gel electrophoresis and then cleaved by a protease such as trypsin. Next, the cleaved peptides are measured with a mass spectrometer. The masses of the measured peptides are compared with the masses of theoretical peptides that were theoretically cleaved from known protein sequences with a protease. The PMF essentially uses a theoretical database containing theoretical peptide fragments of larger proteins and a scoring algorithm to prioritize the candidates.

Several scoring algorithms have been used to identify interesting proteins with PMF. The MOWSE algorithm [18] considers the frequency of peptides against the molecular weights of intact proteins, MASCOT [19] uses a probability-derived score based on the MOWSE algorithm, PeptideProphet uses an algorithm based on Bayesian statistics [20], and Wool Smilansky approach uses binomial statistics [21]. Our theoretical database contains a large amount of protein sequences with which to identify possible alter-

☆ Availability: ASPMF is freely available at http://genepool.kribb.re.kr/ASPMF/.

* Corresponding author.
E-mail address: hurlee@kribb.re.kr (C.-G. Hur).

0006-291X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2008.09.115

native splicing isoforms. Thirty-seven percent of genes within our theoretical database have splicing events, and the protein isoforms from these genes share identical fragment peptides because isoforms produced from a gene share identical exons. In addition, 4.4% of the protein isoforms of our theoretical database contain identical peptides when the proteins are theoretically digested with trypsin. Considering these two factors is important in identifying the true isoforms. However previous algorithms did not consider that protein isoforms produced from a gene share many identical peptides when digested by the same enzyme.

The theoretical database was collected from H-DBAS, TISA, and ECgene, with a total number of 190,529 collected, non-duplicated isoforms. A scoring algorithm was developed to distinguish a true isoform from similar isoforms originating from the identical gene.

Materials and methods

Construction of a theoretical peptide database. There are several published alternative splicing databases, such as TISA, ECgene and H-DBAS, which were constructed using predictions by various statistical and biological data. We combined these databases into a single, theoretical database. Next, the protein sequences from the international protein index (IPI) [22], which widely covers protein sequences assembled from Swiss-Prot, TrEMBL, RefSeq, Ensemble, etc., were added. One of the duplicated transcripts is eliminated in the theoretical databases; the 190,529 transcripts of our theoretical database consisted of 138,523 (51%) transcripts from ECgene, 14,564 (14%) transcripts from H-DBAS and 37,442 (35%) transcripts from IPI. This theoretical database contains both predicted and known protein sequences. The collected protein sequences were theoretically digested by trypsin, and only the fragmented peptides between 500 and 3500 Da were saved in the theoretical database.

The peptides obtained by theoretically digesting the proteins with trypsin were added to the theoretical database as theoretical peptides, leading to a total number of 13,835,840 human peptides. The theoretical database also included protein- and peptide-related information, such as molecular weight and post-translational modification sites [23]. In addition there is also a theoretical peptide database that contains alternative splicing isoforms and tissue-specific information for Arabidopsis. The Arabidopsis theoretical database contains 85,729 isoforms of 26,423 genes, and 3068 genes have alternative splicing events. The 3068 genes generate 30,195 possible isoforms. The isoforms of the two theoretical databases contain information such as missed cleavage sites and post-translational modifications. As shown in Table 1, more than one protein isoform in each gene arose from 37% of the genes in our human theoretical database. These isoforms partially have the same protein sequence because many common exons are shared; however, they differ in sequence length and include several or many different peptide sequences because of frameshift occurrences and the presence of different exons.

Table 1
(A) The number of identical molecular weighted isoforms and (B) the number of protein isoforms

# Identical peptide	# Isoforms (A)	# AS isoforms	# Genes (B)
1<	8334 (4.4 %)	1<	20046 (37%)
2<	3680 (1.9%)	2<	14553 (27%)
3<	2641 (1.4%)	3<	11167 (21%)
4<	2184 (1.2%)	4<	8540 (16%)
5	1298 (0.7%)	5<	6422 (12%)
10<	1035 (0.5%)	10<	2135 (4%)
Max	382 Peptides	Max	170 Isoforms

(A) About 4.4% of isoforms have the peptide fragments of theoretically identical weight. One of them has 382 identical masses. (B) More than 37% of genes have multiple isoforms. One of these genes generates 170 isoforms.

Scoring algorithm. The measured masses were calculated from the mass spectrometry experiments. Matched masses $\{m_1, m_2, \ldots, m_k\}$ were measured masses confirmed against a theoretical database. When searching against the theoretical database, we used the options of tolerance (e.g., 0.3 Da) and coverage. The coverage is how much the peptides match the masses that correlate with a protein isoform. When all the measured peptides are searched among the tolerances against a theoretical database, T is the number of the peptides found in the theoretical database for all the measured masses. Each measured mass is searched between tolerances against the theoretical database. The number of each measured mass $\{f_1, f_2, \ldots, f_k\}$ is divided by T and is denoted by $\{r_1, r_2, \ldots, r_n\}$. If some of the measured masses are not found in the theoretical database, the masses are not considered (e.g., measured peptides O and P in Fig. 1) for scoring. For example, r_1 is the ratio found by dividing f_1 by T. The coverage is calculated by dividing the number of masses matched by the measured masses by the number of all theoretically possible peptides in a protein as a candidate. The coverage is used to limit candidates. The ASPMF algorithm ranks the candidates from low to high values by the P_s score. Here, E_i uses a logarithm to make a lengthy numerical calculation easier to perform. The coverage value is used to prioritize candidates scored identically and filter candidates below a coverage cutoff before scoring them.

$$r_i = f_i/T$$
$$E_i = \log_{10}(r_i)$$
$$P_s = \sum_{i=1}^{k} E_i$$

Results

To evaluate the ASPMF algorithm, we compared it with the MOWSE and binomial statistics-based methods. We used three types of test sets (1) isoforms of genes containing multiple isoforms, (2) isoforms of genes containing a single isoform, and (3) isoforms containing duplicated peptides within each isoform. The masses of the theoretical peptides are extracted from these isoforms in our theoretical human database under each condition.

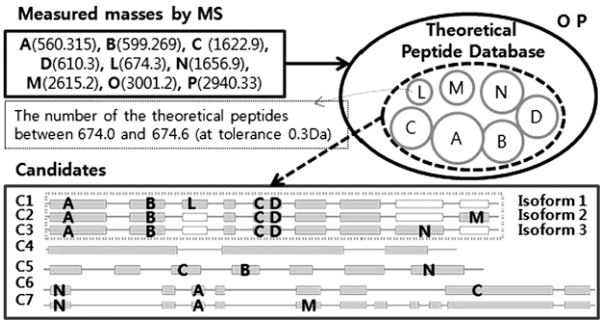

Fig. 1. The ASPMF algorithm scoring process. The dotted circle (T) contains the peptides that are searched against a theoretical peptide database by the measured masses with tolerance ranges (e.g., 0.3 Da). Each gray circle indicates the amount of the peptides searched against the theoretical databases by the measured masses. C1, C2, ..., and C7 are candidates. O and P are not considered in calculating a score (P_s). The dotted rectangle shows the isoforms generated from a gene. C1, C2 and C3 are matched by A, B, C and D in common and are matched to different positions by L, M and N. The circles in the theoretical peptide database are different in size. For example, there are common A, B, C and D and different L, M and N isoforms from a gene. The common masses do not affect distinguishing isoforms. The priority of candidates is decided by L, M and N. Because L is smaller than either M or N, it is the top-ranked isoform candidate (C1).

Performance Evaluation

To evaluate the ability of distinguishing between isoforms, genes having more than five isoforms were extracted from the 189,478 protein sequences in humans. Forty percent of all theoretical peptides of each protein were extracted as true peptides. Less than three peptides with missed cleavage sites were added to the list of true peptides. The false peptides were generated from random values between 500 and 3500 Da. A test set contains an equal number of true and false peptides. The test sets consisted of the true and false sets and were searched against the theoretical database with the options: missed cleavage = 1, tolerance = 0.3 Da, and coverage = more than 10%. The coverage is used to filter worthless candidates. The ASPMF algorithm was compared with the binomial statistics-based method and the MOWSE algorithm. As shown in Fig. 2A, the ASPMF algorithm identified 65.93% isoforms with rank 1 and an average of rank 2.13, where the rank is the order of scores (e.g., the highest score is rank 1) These results indicate that the ASPMF algorithm is better and more stable than the other algorithms (Table 2).

To evaluate the ability of identifying the genes having a single isoform, we tested 504 protein sequences (Table 3). The procedure for making the test sets and the search conditions were identical to those described above. Based on the results within the top 5 score candidates (Table 3 and Fig. 2B), the ASPMF algorithm is better than two others.

There are theoretical peptides with identical molecular weights in many proteins within the theoretical database. As shown in Table 1A, the maximum number of identical peptides in each isoform is 382. To evaluate the algorithms with this condition, we extracted isoforms that have more than five duplicated peptides. The test set consisted of the masses of the true and false peptides. Forty percent of all theoretical peptides of each protein were extracted as true peptides. Duplicated peptides and less than three peptides with missed cleavage sites were added to the list of true peptides. The masses of false peptides were the masses randomly generated between 500 and 3500 Da. A test set contains an equal number of true and false peptides. As shown in Table 4, the ASPMF algorithm improved the accuracy by more than 10%.

System construction

Generating this tool required three main processes: constructing the theoretical database, creating the search algorithm, and developing a user interface. We collected transcripts that are not duplicated in public alternative splicing databases and stored them in the theoretical peptide database. Two scoring methods, MOWSE and ASPMF, were used to determine the priority of the candidate proteins according to the score based on the matched mass list. Thirty-two types of post-translational modifications, molecular weight, and pI are optional user inputs employed to generate the ASPMF interface. The identified isoforms are connected with the Arabidopsis tissue-specific database in TISA and an original website, such as ECgene, H-DBAS, and IPI, by a hyperlink on the web. This tool was developed in PHP and MySQL on Linux. The theoretical pI of each polypeptide is calculated using the "iep" program in the EMBOSS package [24]. As shown in Fig. 3, we used masses as inputs and then searched the masses against a theoretical data-

Table 2
Test results for genes that have multiple isoforms

Algorithm	MOWSE	Binomial	ASPMF
Test set	4226	4226	4226
Avg. rank	3.58	3.53	2.13
Rank 1	2435 (57.62%)	1469 (34.76%)	3268 (77.33%)
Rank 2	709 (16.78%)	972 (23.0%)	374 (8.85%)
Rank 3	316 (7.48%)	647 (15.31%)	190 (4.5%)
Rank 4	154 (3.64%)	381 (9.0%)	89 (2.1%)
Rank 5	117 (2.77%)	230 (5.44%)	67 (1.59%)
Total	3731 (88.29%)	3699 (87.53%)	3988 (94.37%)

Isoforms having more than five isoforms (Table 1B). The 'test set' is the number of test sets. The figures in parenthesis are the ratios between the matched count and the number of the test sets. The 'Avg. rank' is the average of all ranks. This result is searched and scored by options-tolerance (< 0.3 Da) and missed cleavage (0 and 1).

Table 3
Test results for genes that have a single isoform

Algorithm	MOWSE	Binomial	ASPMF
Test set	504	504	504
Avg. rank	1.53	2.6	1.68
Rank 1	425 (84.33%)	245 (48.61%)	433 (85.91%)
Rank 2	23 (4.56%)	117 (23.14%)	24 (4.76%)
Rank 3	11 (2.18%)	52 (10.32%)	15 (2.98%)
Rank 4	11 (2.18%)	35 (6.94%)	8 (1.39%)
Rank 5	1 (0.2%)	15 (2.98%)	5 (0.99%)
Total	471 (93.45%)	464 (93.45%)	485 (96.23%)

This result is scored by option-tolerance (< 0.3 Da) and missed cleavage (0 and 1).

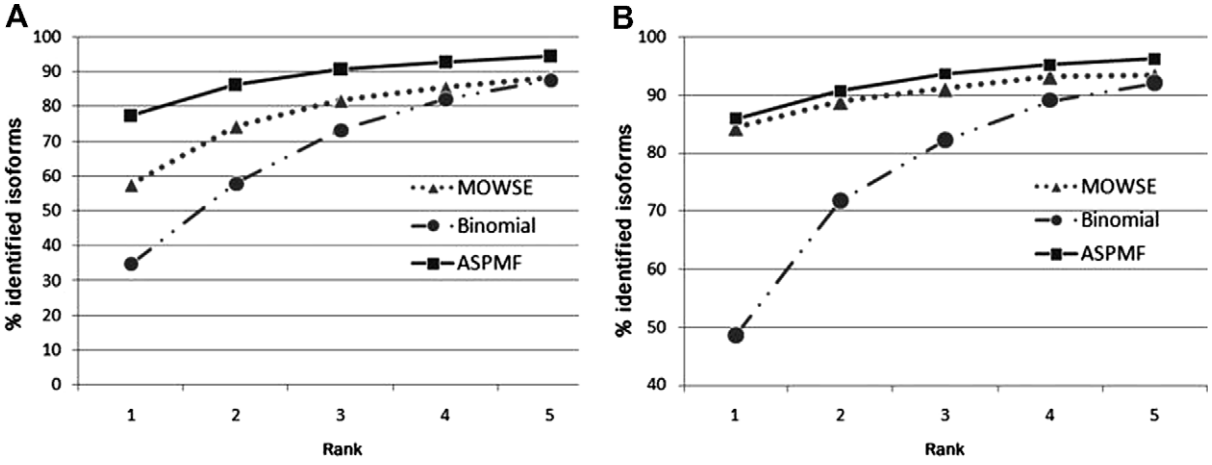

Fig. 2. Performance comparison among the three algorithms. There are two results above (A) the test results for genes that have multiple isoforms (detailed information in Table 2) and (B) the test results for genes that have a single isoform (detailed information in Table 3) in a human theoretical database considering alternative splicing. All tests are evaluated with options (tolerance: +/− 0.3 Da and missed cleavage: up to 1). The figures are the accumulation graphs that were drawn by accumulating the percentage value of each rank.

Table 4
Test results of isoforms that have peptide fragments with identical weights

Algorithm	Binomial	ASPMF
Test set	1310	1310
Avg. rank	6.08	3.09
Rank 1	670 (51.15%)	844 (64.43%)
Rank 2	194 (14.81%)	200 (15.27%)
Rank 3	94 (7.18%)	93 (7.10%)
Rank 4	70 (5.34%)	52 (3.97%)
Rank 5	41 (3.13%)	23 9 (1.78%)
Total	1069 (81.6)	1212 (92.52%)

The isoforms of duplicated peptides were evaluated. This result is scored by option-tolerance (< 0.3 Da) and missed cleavage (0 and 1).

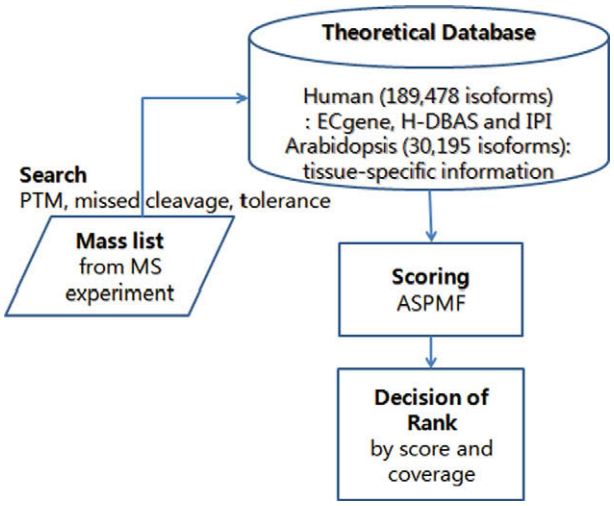

Fig. 3. The system process. This system receives PTM, missed cleavage, tolerance and, etc. as input. With respect to these inputs, masses are searched against theoretical database. Isoforms less than 10% in coverage are eliminated. Remaining isoforms are scored by ASPMF algorithm, and then candidates of the same score are ordered by coverage.

base. The search results are filtered by coverage first and then used to calculate the score. If identical scores result, the candidates are prioritized by coverage.

Conclusion

Alternative splicing is now largely accepted as one of the mechanisms that explains the discrepancy between the high-level complexity of humans and their relatively small number of genes. A gene can have different transcripts that have unique functions. There have been strong efforts towards discovering alternative splicing isoforms and determining their characteristics. In order to detect and identify alternative splicing isoforms, we used PMF with a theoretical database and scoring algorithm. The theoretical database properly consolidated the public alternative splicing databases based on bioinformatic and biological data. A gene with various numbers of isoforms is apt to generate identical peptides after enzymatic digestion, making it difficult to identify isoforms. Our results showed that the ASPMF algorithm is efficient in identifying alternative splicing isoforms; moreover, additional information, such as protein molecular weight, pI, and so on, and can make isoform identification more accurate. Our online program will help researchers identify alternative splicing isoforms in an easily accessible manner.

Acknowledgments

This work was supported by grants from the Crop Functional Genomics Center (CFGC) and Microbial Genomics & Applications Center (MGAC), one of the 21st Century Frontier Research Programs of the Ministry of Education, Science, and Technology (MEST).

References

[1] B. Modrek, C. Lee, A genomic view of alternative splicing, Nat. Genet. 30 (2002) 13–19.
[2] M. Krawczak, N.S.T. Thomas, B. Hundrieser, M. Mort, M. Wittig, J. Hampe, D.N. Cooper, Single base-pair substitutions in exon-intron junctions of human genes: nature distribution and consequences for mRNA splicing, Hum. Mutat. 28 (2007) 150–158.
[3] J. Hull, S. Campino, K. Rowlands, M. Chan, R.R. Copley, M.S. Taylor, K. Rockett, G. Elvidge, B. Keating, J. Knight, D. Kwiatkowski, Identification of common genetic variation that modulates alternative splicing, PLoS 3 (2007) 1009–1018.
[4] N.A. Faustino, T.A. Cooper, Pre-mRNA splicing, human disease, Genes Dev. 17 (2003) 419–437.
[5] J.X. Zhu, E. Dagostino, P.A. Rejto, B. Mroczkowski, B. Murray, Identification and characterization of a novel and functional murine Pin1 isoform, Biochem. Biophys. Res. Commun. 3 (2007) 529–535.
[6] S. Saito, N. Takahashi-Sasaki, W. Araki, Identification characterization of a novel human APH-1b splice variant lacking exon 4, Biochem. Biophys. Res. Commun. 4 (2005) 1068–1072.
[7] C.C. Wang, Q. Zeng, L.A. Hwang, K. Guo, J. Li, H.C.Liew.W. Hong, Mouse lymphomas caused by an intron-splicing donor site deletion of the FasL gene, Biochem. Biophys. Res. Commun. 1 (2006) 50–58.
[8] S. Noh, K. Lee, H. Paik, C. Hur, TISA: tissue-specific alternative splicing in human and mouse genes, DNA Res. 13 (2006) 229–243.
[9] Y. Lee, Y. Lee, B. Kim, Y. Shin, S. Nam, P. Kim, N. Kim, W.H. Chung, J. Kim, S. Lee, ECgene: an alternative splicing database update, Nucleic Acids Res. 35 (2007).
[10] J. Takeda, Y. Suzuki, M. Nakao, T. Kuroda, S. Sugano, T. Gojobori, T. Imanishi, H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-invitational, Nucleic Acids Res. 35 (2007) D104–D109.
[11] H. Ner-Gaon, N. Leviatan, E. Rubin, R. Fluhr, Comparative cross-species alternative splicing in plants, Plant Physiol. 144 (2007) 1632–1641.
[12] B. Wang, V. Brendel, Genome wide comparative analysis of alternative splicing in plants, Proc. Natl. Acad. Sci. USA 103 (2006) 7175–7180.
[13] S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigo, S.P. Briggs, V. Bafna, Improving gene annotation using peptide mass spectrometry, Genome Res. 17 (2007) 231–239.
[14] M.C. Giddings, A.A. Shah, R. Gesteland, B. Moore, Genome-based peptide fingerprint scanning, Proc. Natl. Acad. Sci. USA 100 (2003) 20–25.
[15] K. Nagao, N. Togawa, K. Fujii, H. Uchikawa, Y. Kohno, M. Yamada, T. Miyashita, Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays, Hum. Mol. Genet. 14 (2005) 3379–3388.
[16] D.D. Shoemaker, E.E. Schadt, C.D. Armour, Y.D. He, P. Garrett-Engele, P.D. McDonagh, P.M. Loerch, A. Leonardson, P.Y. Lum, G. Cavet, L.F. Wu, S.J. Altschuler, S. Edwards, J. King, J.S. Altschuler, S. Edwards, J. King, J.S. Tsang, G. Schimmack, J.M. Schelter, J. Koch, M. Ziman, M.J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M.R. Meyer, M. mao, J. Burchard, M.J. Kidd, H. Dai, J.W. Phillips, P.S. Linsley, R. Stoughton, S. Scherer, M.S. Boguski, Experimental annotation of the human genome using microarray technology, Nature 409 (2001) 922–927.
[17] T.A. Clark, A.C. Schweitzer, T.X. Staples, G. Lu, H. Wang, A. Williams, J.E. Blume, Discovery of tissue-specific exons using comprehensive human exon microarrays, GenomeBiology.com 8 (2007).
[18] D.J. Pappin, P. Hojrup, A.J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biol. 3 (1993) 327–332.
[19] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (1999) 3551–3567.
[20] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (2002) 5383–5392.
[21] A. Wool, Z. Smilansky, Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting, Proteomics 2 (2002) 1365–1373.
[22] P.J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, R. Apweiler, The international protein index: an integrated database for proteomics experiments, Proteomics 4 (2004) 1985–1988.
[23] N. Blom, T. sicheritz-Ponten, R. Gupta, S. Gammeltoft, S. Brunak, Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence, Proteomics 4 (2004) 1633–1649.
[24] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet. 16 (6) (2000) 276–277.