research articles Journal of proteome research

Detection of Alternative Splice Variants at the Proteome Level in *Aspergillus flavus*

Kung-Yen Chang,[†,§] D. Ryan Georgianna,[‡,||] Steffen Heber,[†] Gary A. Payne,[‡] and David C. Muddiman*[,§]

Bioinformatics Research Center, Center for Integrated Fungal Research, and W.M. Keck FT-ICR-MS Laboratory, Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695, and Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710

Received July 9, 2009

Identification of proteins from proteolytic peptides or intact proteins plays an essential role in proteomics. Researchers use search engines to match the acquired peptide sequences to the target proteins. However, search engines depend on protein databases to provide candidates for consideration. Alternative splicing (AS), the mechanism where the exon of pre-mRNAs can be spliced and rearranged to generate distinct mRNA and therefore protein variants, enable higher eukaryotic organisms, with only a limited number of genes, to have the requisite complexity and diversity at the proteome level. Multiple alternative isoforms from one gene often share common segments of sequences. However, many protein databases only include a limited number of isoforms to keep minimal redundancy. As a result, the database search might not identify a target protein even with high quality tandem MS data and accurate intact precursor ion mass. We computationally predicted an exhaustive list of putative isoforms of *Aspergillus flavus* proteins from 20 371 expressed sequence tags to investigate whether an alternative splicing protein database can assign a greater proportion of mass spectrometry data. The newly constructed AS database provided 9807 new alternatively spliced variants in addition to 12 832 previously annotated proteins. The searches of the existing tandem MS spectra data set using the AS database identified 29 new proteins encoded by 26 genes. Nine fungal genes appeared to have multiple protein isoforms. In addition to the discovery of splice variants, AS database also showed potential to improve genome annotation. In summary, the introduction of an alternative splicing database helps identify more proteins and unveils more information about a proteome.

Keywords: mass spectrometry • proteomics • alternative splicing • database • *Aspergillus flavus*

Introduction

Protein identification is the key part of mass spectrometry-based proteomic analysis. In a typical "Bottom-Up" approach, proteins of interest are separated by gel electrophoresis, digested by a specific enzyme, and followed by analysis of single- or multidimensional chromatography coupled with tandem mass spectrometry. The acquired mass spectral data are then used to search against databases to find matched proteins. In practice, a major portion of acquired spectra find no matched sequence entry in protein databases.[1] Probable explanations include quality of spectrum, search engine algorithm, sequence polymorphism, post-translational modification, and transcriptional variation by RNA splicing and editing.[1,2]

Alternative pre-mRNA splicing is a mechanism that removes the intervening introns (noncoding sequences) and joins the flanking exons (coding regions) in different arrangements. This process allows a single gene to generate various mRNAs, then multiple protein variants, which might have diverse and even antagonistic functions. Alternative splicing has been found to play important roles in many cellular and developmental processes in metazoans.[3−5] Aberrant splicing has been implicated in various diseases including cancer.[6] Analyses of expressed sequence tags (ESTs) and alternative splicing microarray data estimated that more than two-thirds of human genes are alternatively spliced.[7] As a demonstration of diversity achieved by alternative splicing, the *Drosophila Dscam* gene has the potential to encode 38 016 distinct spliced variants, nearly 3 times the total number of genes in *Drosophila*.[8] Although fungi appear to use alternative splicing less frequently than metazoans, a genome-wide survey in *Cryptococcus neoformans* Serotype D, a basidiomycete yeast found ubiquitously in the environment, revealed evidence of alternative splicing for 277 genes or 4.2% of the total.[9] These data suggest that alternative splicing in fungi is more prevalent than previously thought.

* Author for Correspondence: David C. Muddiman, Ph.D., W.M. Keck FT-ICR Mass Spectrometry Laboratory, Department of Chemistry, North Carolina State University, Raleigh, North Carolina 27695. Phone: 919-513-0084. Fax: 919-513-7993. E-mail: david_muddiman@ncsu.edu.

† Bioinformatics Research Center, North Carolina State University.

§ W.M. Keck FT-ICR-MS Laboratory, Department of Chemistry, North Carolina State University.

‡ Center for Integrated Fungal Research, North Carolina State University.

|| Duke University Medical Center.

10.1021/pr900602d

Published on Web 01/04/2010

One of the challenges in mass spectrometry-based proteomics is that tandem MS peptides can only identify the proteins if the target database contains the correct sequences for peptide mapping. Today's protein databases vary in many aspects such as level of curated annotation, size of records and species, and degree of sequence redundancy. Currently, only a small percentage of estimated alternative splicing variants are deposited in the major protein databases. Databases like Swiss-Prot[10] and RefSeq[11] were originally designed to keep a minimum level of redundancy. Among 408 099 sequence entries in Swiss-Prot (release 14.7, Jan 20, 2009), only 27 169 additional sequences (6.6%) are produced by alternative splicing, initiation or promoter usage, or ribosomal frameshifting. Considering the possibility that true protein isoforms are actually absent from the target database, the search will not be able to identify the correct protein even with high quality tandem mass data and accurate intact precursor ion mass.

Alternatively spliced variants can be detected by alignment comparison of transcripts and genomic sequences and data mining of the scientific literature. Recently, high-throughput tandem mass spectra were utilized to find novel splice variants of previously annotated genes.[12] The peptide sequences derived from tandem mass spectra provide evidence of translation products. All popular search algorithms employed in protein identification pipelines rely on a protein database to infer peptide sequences from MS/MS spectra. An alternative to the database search is *de novo* sequencing, which extracts the peptide sequence directly from the spectrum without the help of any sequence database. The hybrid approaches[13−15] which infer short sequence tags (partial sequences) by *de novo* sequencing, followed by a database search using these tags to find the peptides in the sequence database, may be helpful for the analysis of post-translationally modified peptides and splice variants.

Many groups independently developed alternative splicing databases which focus on detection and storage of AS sequences, such as ASG,[16] HOLLYWOOD,[17] ECGene,[18] ASAP II,[19] ASTD,[20] and LSAT.[21] The detected or generated alternatively spliced isoforms may vary between different databases due to different input data, genome assembly, method for alignment and level of stringency.[19] Most AS databases above are not actively updated on a regular schedule. In addition, many AS databases only provide limited query and download functions and focus on selected species. These limitations hinder the utilization of existing AS databases in proteomic studies.

Aspergillus flavus is a filamentous ascomycete fungus that is able to infect economically important crops, such as maize, cotton, tree nuts, and peanuts, while contaminating them with potent mycotoxins.[22] Among the many secondary metabolites produced by *A. flavus*, aflatoxin B1 is the most toxic and potent hepatocarcinogenic natural compound ever characterized.[23] Consumption of aflatoxins can cause liver damage including acute hepatitis, immunosuppression, and hepatocellular carcinoma.[24] The primary assembly of the *A. flavus* genome indicates that it consists of eight chromosomes and is 36.3 Mb in size. The genome contains 13 071 predicted genes and the mean gene length is 1384 bp.[25] The bottom-up and top-down proteomic profiles of *A. flavus* under different temperatures have already been surveyed through a SILAC approach.[26,27] It has been reported that intron splicing is essential for amine-regulated gene expression in *Aspergillus oryzae*,[28] a widely used industrial and food fungus which is almost genetically identical to *A. flavus*. Combined with the biological and economic importance, established knowledge of its genome and proteome, and existence of splicing events in filamentous fungi, *A. flavus* proves to be an effective model organism.

Without an AS database of *A. flavus* available, we built our own AS database to test two hypotheses: (1) that we can confidently identify more proteins using an AS database combined with accurate mass precursor and tandem-MS data; and (2) that fungus undergoes sufficient alternative splicing that it can be detected at the proteome level. The overview of our study is illustrated in Figure 1.

Experimental Procedures

Original Protein Database of *A. flavus*. The whole genome sequencing of *A. flavus* strain NRRL 3357, which provided 5-fold sequence coverage, was carried out at The J. Craig Venter Institute (JCVI, formerly TIGR). Automated annotation was also conducted at JCVI and additional manual annotation was coordinated through North Carolina State University (NCSU).[29] A collection of 12 832 annotated protein sequences acquired from The Center for Integrated Fungal Research at NCSU is referred as the 'original database' in the remainder of the article.

Construction of Alternative Splicing Database. Since the original database did not include any prediction of potential alternative isoforms, we applied the approach of aligning cDNAs with genomic sequences to construct an alternative splicing database of *A. flavus*. A total of 20 371 ESTs were downloaded from the EST database of NCBI with a species filter specifying '*Aspergillus flavus*'. The sequences of 12 832 predicted genes were acquired from The Center for Integrated Fungal Research at NCSU. Instead of directly aligning all ESTs to genomic sequences, we first used BLAST[30] to select those having similar sequences with the previously annotated *A. flavus* genes. Each EST was only allowed to be assigned to one gene. If an EST was similar to several different genes, it was assigned to the top ranked hit. The cutoff expect value in BLAST mapping was 0.001. The screening mapped 16 121 ESTs to 4497 genes.

To consider possible exons located in upstream and downstream regions of originally annotated genes, we extended both 5′ and 3′ ends of genes by 3 kb which equaled to twice the length of the longest intron found in the original database. The gene transcript derived from the previous annotation was considered as another EST to the gene. The alignments of transcript and ESTs to genomic sequences were performed using the *sim4* program,[31] which unveiled the boundaries of exons and introns. The large amount of information carried in ESTs was merged and integrated into a single splicing graph.[32] In the splicing graph, paths represented transcripts and vertex with multiple incoming or outgoing edges corresponded to alternative splice sites. Alternative splicing events were detected as bifurcations in the graph (Figure 2A). This procedure might not identify truncated transcripts.[16] While comparing genomic positions of boundaries belonging to the same exon, a 10 bp allowance was made. If two or more 5′ or 3′ splice sites of the same exon were located within 10 bp, the predominant one was kept. Otherwise, they were considered alternative splice sites. It should be noted that this procedure might result in losing potential variants. The updated information and detected alternative splicing events reconstructed the gene model. In our graphic representation of gene structure, nodes and edges represented exons and introns, respectively (Figure 2C). All putative protein isoforms of a gene were predicted by visiting

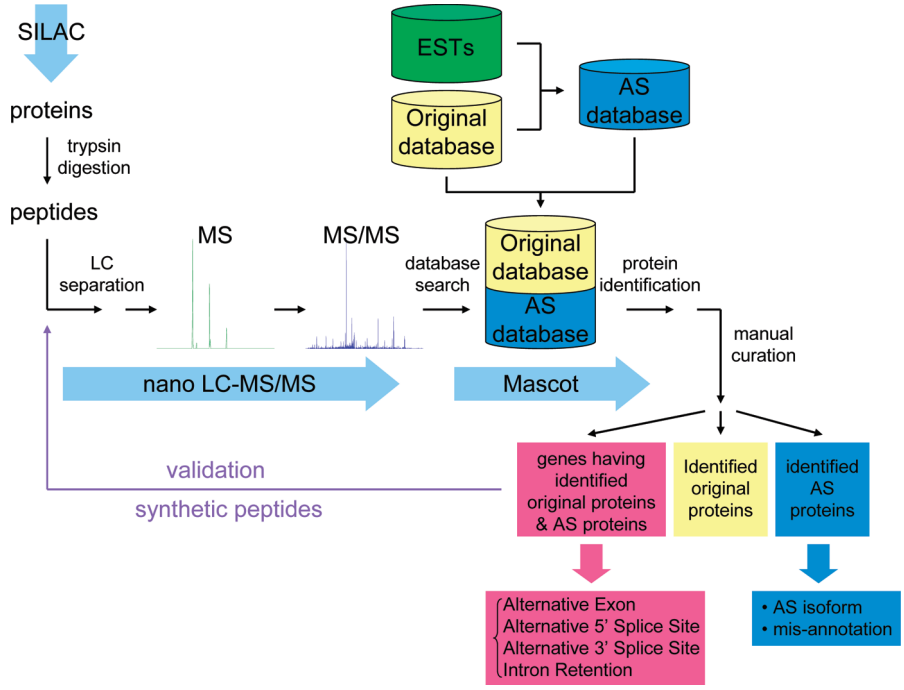

Figure 1. Integration of alternative splicing database in bottom-up proteomic analysis. Tandem mass spectra with high mass accuracy were generated from SILAC experiment followed by nano LC-MS/MS analysis. By including predictions of alternative splicing variants in database search, new protein isoforms were detected. Endogenous peptide identifications were validated by synthetic peptides.

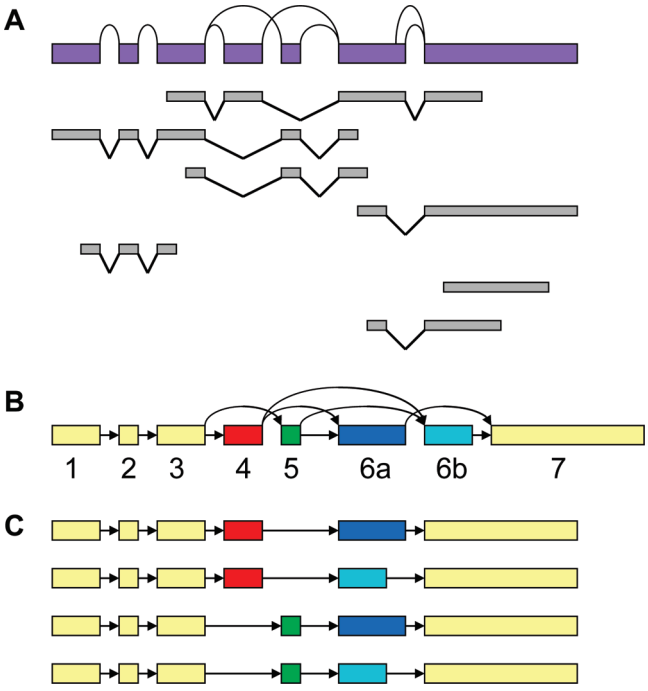

Figure 2. Detection of alternative splicing and generation of putative alternative splicing variants. (A) Visualization of splicing graph (purple). The example gene consisted of 7 exons based on the alignments of ESTs (gray). Two types of alternative splicing were detected using splicing graph. Exon 4 and 5 were mutually exclusive exons. Exon 6 had two alternative 3′ splice sites. (B) Directed acyclic graph (DAG) representation of gene structure. (C) Generation of all potential putative transcripts.

all paths in the graph, and followed by generating the corresponding sequences (Figure 2D).

Since the exact translation initiation site of a gene was often unclear, 3-frame translations (forward) were also considered in our prediction. The protein products of genes were translated according to the standard code. The database only kept protein sequences which initiated with a start codon, ended with a stop codon, and had at least 18 amino acids (the shortest sequence found in the original protein database). The final step of construction was the elimination of redundancy. We removed any newly predicted sequence which was either subsequence or fully identical with one in the original database.

SILAC Experiment. We used the existing mass spectra to test the newly constructed alternative splicing database. The original purpose of the experiments was to study the relative change of protein levels between conducive (28 °C) and nonconducive (37 °C) temperatures for aflatoxin biosynthesis. The experiment was previously described in full detail.[26] In short, different labeled cultures of *A. flavus* were grown for 24 h at 28 or 37 °C. Extracted protein samples were separated on 12.5% SDS-PAGE gel. Forty bands were sliced from each lane. Samples were reduced, alkylated, and followed by trypsin digestion for 18 h at 37 °C. Then each in-gel digested samples was analyzed by nanoflow LC-MS/MS on a LTQ-FT (ThermoFisher Scientific).

Database Search. To study the effects of the new AS database on protein identification, the newly predicted alternatively spliced variants were combined with the originally annotated proteins. The existing set of experimental spectra as described previously[26] was searched against the combined database. The protein search algorithm used in the analysis was Mascot Server version 2.2.04 (Matrix Science Ltd.). The search allowed two missed cleavage sites, 5 ppm peptide tolerance, 1 Da MS/MS fragment ion tolerance, two variable modifications, Deamidated (NQ) and Oxidation (M), and one fixed Cys modification, Carbamidomethyl (C).

Mascot incorporates a probability based implementation of the Mowse scoring algorithm.[33] The Mascot score is reported as $-10\ \mathrm{Log}_{10}(P)$, where P is the absolute probability that the

observed match is a random event. The absolute probability P is equal to $E \times N^{-1}$, where E is the expect value and N is the number of proteins in the database. The significance threshold for the search was 0.05 ($p < 0.05$). An event is significant if it would be expected to occur at random with a frequency of less than 5%. Knowing the size of the combined database, the corresponding cutoff Mascot score would be $(-10)\ \log[(1/(12832 + 9807)) \times 0.05] = 56.56$.

Decoy Database Search Strategy. Controlling the false discovery rate (FDR) is a commonly accepted approach to multiple testing correction in large-scale mass spectrometry-based proteomic studies. FDR represents the percentage of significant peptide−spectrum matches (PSMs) and can be estimated using the target-decoy strategy. The 'target' database was the expanded *A. flavus* protein database. A corresponding 'decoy' database was created by reversing the sequences in the target database. Two separate searches were performed against the target and decoy databases individually using identical search parameters. At the significance threshold of 0.05 ($p < 0.05$) in Mascot searches, 14 615 target PSMs and 407 decoy PSMs above the threshold were counted. A FDR of 2.78% was estimated by computing the ratio of decoys and targets.[34] It means that if we accept 100 tandem mass spectrum assignments, then we expect less than three of those identifications to be incorrect.

Validation of AS Peptides. Previous work from our laboratories already showed that synthesized peptides can be used to validate the identification of naturally processed HLA class II peptides.[35] Because alternative isoforms from the same gene usually share a partial sequence, the peptides which were specific to the AS protein and not found in the original database became vital to the identification of isoforms. We ordered the synthetic peptides of the detected endogenous peptide sequences. The synthetic peptides were analyzed by nLC-MS/MS. The tandem mass spectra from the synthetic peptides were compared to those from the corresponding endogenous peptides.

Data Availability. The Bottom-Up SILAC *A. flavus* .RAW data files can be downloaded from Tranche distributed file system (tranche.proteomecommons.org) by providing the following hash, O9h2YUGGpAOG+ex5+rYTySoRxqvyPayGlWPspibKkA13-BXCVcpVMp3oCmH4HwZOofp5azAQcx4coCH6I82DCx5vQj-wwAAAAAAAAn5g==.[26]

Results

To study the importance of alternative splicing in proteomic analysis, a customized AS database of *A. flavus* was built based on ESTs. The original database consisted of 12 832 previously annotated proteins but no splice isoforms. The newly constructed AS database provided additional 9807 eligible alternatively spliced variants and expanded the size of the original database by 76%. According to our predictions, 1292 *A. flavus* genes had multiple protein isoforms, 8.59 putative isoforms per gene on average.

To evaluate the newly predicted variants, we combined the AS database with the original database. A set of bottom-up mass spectrometry data generated from the previous analysis[26] which studied protein profiles at two growing temperatures of *A. flavus* was used to search against the combined database by Mascot. The primary goal of this study is to investigate whether total protein identifications would increase after introducing the AS database. Although the data set was originally derived from a SILAC experiment, it is beyond the scope of this paper to indentify the splicing variants responding to different environment stimuli. A total of 556 proteins were identified from the original database in the nonconductive and conductive conditions combined. The identifications detected from the AS database were carefully examined by extensive manual curation. Five chosen peptides were subsequently synthesized and analyzed by nLC-MS/MS to validate the identifications.[35] At the end, the Mascot search identified 29 new proteins encoded by 26 *A. flavus* genes from the existing tandem mass spectra (Table 1). Only one new protein was identified by a single peptide. Up to 41 previously unseen MS/MS peptides were included in the matches, 34 of them had EST evidence. For those identified peptides which had no overlapped ESTs, they were mapped to genome sequences considered as transcription regions in previous annotation. All 41 peptide sequences were absent in the original protein database, indicating they were derived from different gene models. The alignments of tandem MS peptides to genomic sequences suggested that the variants were generated by several patterns of alternative splicing, including cassette alternative exons, alternative 5′ splice site, alternative 3′ splice site and intron retention.

Peptide sequences resulting from proteolytic digestion of the sample proteins expressed in higher eukaryotes can be present in multiple protein isoforms. These shared peptides cannot be used to discriminate between different alternative splice variants. A solid conclusion of the detection of alternative isoforms would be possible if different protein products from one gene were identified by individually unique peptides. Nine genes from our results had at least two protein variants supported by different sets of tandem MS peptides. Four patterns of alternative splicing are presented in the following examples.

Cytochromes b5 are ubiquitous electron-transport proteins involved in a variety of biochemical processes and metabolic pathways. NADH-cytochrome b5 reductase (EC 1.6.2.2) serves as electron donor for cytochrome b5.[36] In rat, the NADH-cytochrome b5 reductase gene generates two transcripts by a combination of alternative promoters and alternative initiation of translation: a ubiquitous mRNA coding for the myristylated membrane-bound form, and an erythroid mRNA which generates both the soluble form and a nonmyristylated membrane-binding form.[37] In comparison with the original protein of *A. flavus* NADH-cytochrome b5 reductase, the AS protein contained two additional exons in the middle and ended with a shorter exon (Figure 3). Two unique peptides EAVSGVTIASALLTK (spectrum see Figure 3) and AVLRPYTPTTMK were specific to the AS protein. Not only were both peptides aligned in the intron region of the original protein, but they specified the splice sites of the two alternative exons and proved the translation across the constitutive and alternative exons in the AS protein. The additional exons also caused a frame shift of protein translation and resulted in an early stop site for the AS protein. Four peptides which were aligned to the last exon of the original protein, beyond the stop site of the AS variant, allowed the identification of the original protein. The peptide of the sequence EAVSGVTIASALLTK was synthesized and analyzed by the same LTQ-FT to validate the identification (Figure 3). Because of the higher abundance of synthetic peptide, a stronger signal was observed in the tandem mass spectrometry measurements but the sequence ions of the endogenous and synthetic peptides were consistent.

Homocysteine is an intermediate metabolite of the essential methionine. Cystathionine beta-synthase (EC 4.2.1.22) catalyzes

Table 1. List of Proteins Identified from Alternative Splicing Database

NCBI RefSeq	description	protein score	total peptide matches	peptide sequence[a]	peptide expection value	counts of peptide detection	EST evidence	AS pattern	original protein detection
XP_002374454.1	NADH-cytochrome b5 reductase	183	5	EAVSGVTIASALLTK[b,d]	1.20×10^{-9}	2	Y	Alternative Exon	Y
				AVLRPYTPTTMK	1.3	2	Y		
XP_002374667.1	pyruvate decarboxylase PdcA	512	8	VSVVLTK	0.12	1	Y	Alternative 3′ Splice Site	Y
XP_002374021.1	glyceraldehyde 3-phosphate dehydrogenase GpdA	64	3	TIKAASEEGELKGK	0.69	1	Y	Alternative 3′ Splice Site	Y
XP_002378186.1	cystathionine beta-synthase[c]	72	3	MQLSSALLMK	1.1	1	Y	Alternative 5′ Splice Site, Alternative 3′ Splice Site	Y
				TRTSDLPSTLQPHEQK	0.56	1	Y	Alternative 3′ Splice Site	
		101	4	KSGRRPR	0.9	1	Y	Alternative 5′ Splice Site	
				WPTNWVQRK	3.4	1	Y	Alternative 3′ Splice Site	
XP_002378559.1	myo-inositol-phosphate synthase	236	8	QLQPPGQQRR	3.8	1	Y	Intron Retention, Alternative 3′ Splice Site	Y
XP_002372782.1	mitochondrial F1 ATPase subunit alpha	212	8	RPRLPSR	0.34	1	Y	Intron Retention	Y
XP_002376446.1	sulfide quinone reductase	132	3	STMNSWLWFPALTSTM VALRVCPR	2.6	1	N	Alternative 5′ Splice Site, Intron Retention Intron Retention	Y
XP_002384978.1	actin binding protein[c]	61	4	FGANQSFVGTKPPLPSGSMPKPTSAVAP VGSASRTFADEGGK	6.9	1	Y	Alternative 3′ Splice Site	N
		73	3	TMRLSSGKVNM	0.27	2	Y	Alternative 3′ Splice Site	
		61	3	VANKFGANQSGTR	2.3	1	N		
XP_002382549.1	60S ribosomal protein L32	84	15	AQEGILLR	0.011	15	N	Alternative Exon	N
XP_002385010.1	conserved hypothetical protein	121	2	TMAGLFKGIHGVVQDK	3.7	1	N	Alternative Exon	N
XP_002385145.1	ribosomal protein S13p/S18e	159	6	EDLERLKK	0.97	2	Y	Alternative 3′ Splice Site	N
		71	4	HYWGLRVRGQHTNR	2.5	1	Y		
XP_002375430.1	conserved hypothetical protein	64	12	VMGMKQFPR	0.0009	12	Y	Alternative 5′ Splice Site	N
XP_002375948.1	peroxiredoxin	146	2	VENNDILFLSDPDAK[b]	1.00×10^{-9}	1	Y	Alternative 3′ Splice Site	N
				VSGAEAVLAHL[b]	5.50×10^{-7}	1	Y		
		125	4	ANKVENNDILFLSDPDAK[b]	0.31	1	Y		
XP_002376287.1	transcription elongation complex subunit (Cdc68)	83	34	HNTRCGR	0.0026	34	N	Alternative 5′ Splice Site, Alternative 3′ Splice Site	N
XP_002375258.1	14-3-3 protein sigma, gamma, zeta, beta/alpha	151	3	DNLTLWTSSDGQEPEGAASK[b]	8.30×10^{-9}	2	Y	Alternative Exon	N
				EDKPEEESAPAPEDKGEESKPAAPES	0.36	1	Y		
XP_002374498.1	FAD dependent oxidoreductase	285	19	NGAPIKGLWAAGEVTGGLHGQNR	1.9	3	Y	Intron Retention	N
XP_002374268.1	iron superoxide dismutase A	538	11	TYANQDPVVGQFQPLLGI DAWEHAYYLQYQNR	3.20×10^{-5}	1	Y	Alternative 3′ Splice Site	N
XP_002378098.1	acetyl-CoA carboxylase[c]	82	3	GHTSSMLWSNMSSPSSSTSRPSVNSV VVLGSSLTLQSIQTKWR	1.8	2	N	Alternative Exon	N
				TESVAADVAQLLIGNK	2.80×10^{-8}	1	Y		
		1441	42	EQLSEIK	0.12	1	Y	Alternative 3′ Splice Site	N
				TESVAADVAQLLIGNK	2.40×10^{-8}	3	Y		
XP_002378109.1	conserved hypothetical protein	111	5	EFEDAAFALQPGQVSGI VDTASGVHLIER	2.90×10^{-6}	1	Y	Alternative 3′ Splice Site	N
				SKEEAIEILR	0.00013	4	Y		
XP_002380658.1	UTP-glucose-1-phosphate uridylyltransferase Ugp1	444	11	APATETSNAGSFGK	2.00×10^{-7}	1	Y	Alternative Exon	N
		125	4	AAKALPTHLRAPATETSNAGSFGK	3.6	1	Y		
XP_002373453.1	prefoldin subunit 6	69	2	AEILQYQSQMQQQAAAASASA	2.80×10^{-6}	2	Y	Alternative 3′ Splice Site	N
XP_002377133.1	mitochondrial acetolactate synthase small subunit	115	3	LIAPFGVLESTR[d]	6.60×10^{-8}	2	Y	Alternative 3′ Splice Site	N
				VLDISNNNCIVEVSAKPSRIDSFMK[d]	5	1	Y		
XP_002377088.1	cyclophilin	192	7	SGELESEDKGSHEEL	9.40×10^{-8}	6	Y	Alternative 3′ Splice Site	N
XP_002384637.1	phosphofructokinase	61	1	NDQTSTIYSTTEIANIIK	3.90×10^{-6}	1	Y	Alternative 3′ Splice Site	N
XP_002379270.1	hypothetical protein	64	9	NRILNNNKK	0.00071	9	N	Alternative Exon	N
XP_002379566.1	conserved hypothetical protein	85	3	LLLALVSK[d]	0.0085	1	Y	Alternative 3′ Splice Site	N

[a] Only peptides detected from the AS database search are listed. [b] Peptides which were validated by synthetic peptides. [c] Genes which had the identifications of multiple different isoforms. [d] Peptides which were included in the latest gene annotation.

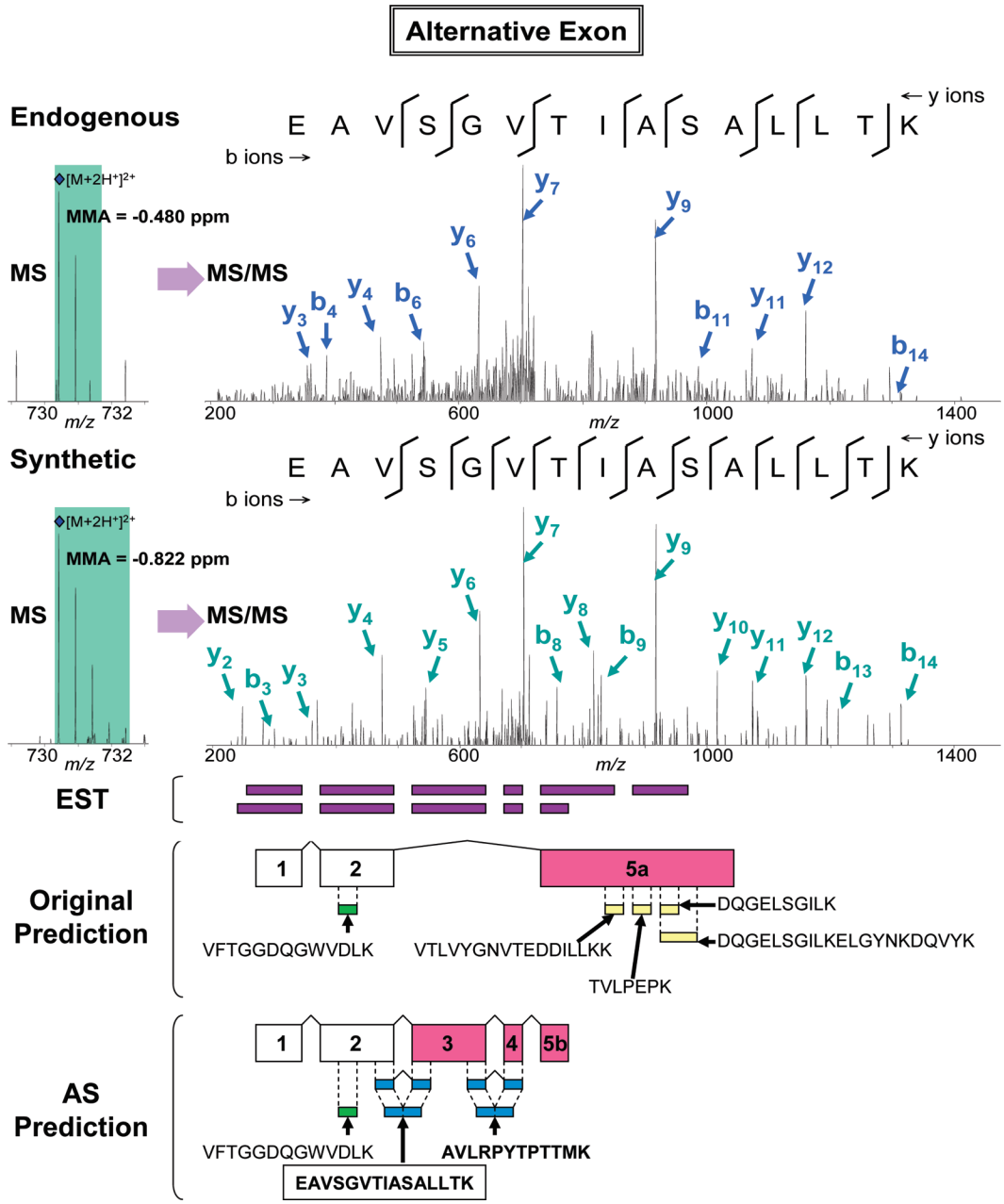

Figure 3. NADH-cytochrome b5 reductase. Two alternative exons appeared in the middle of the AS protein (lower trace). The MS/MS spectrum of peptide EAVSGVTIASALLTK (upper trace) was resulted from a precursor ion scan at *m/z* 730.42 with a measured mass accuracy of 0.48 ppm. The tandem MS spectrum of a synthesized counterpart (middle trace) was resulted from a precursor ion scan at the same *m/z* with a mass accuracy of 0.82 ppm. The AS protein-specific peptide proved the junction of the constitutive and alternative exons. Original protein-specific, AS protein-specific, and indistinguishable shared peptide were labeled in yellow, blue and green, respectively.

the conversion of homocysteine to cystathionine, the first step in the transsulfuration pathway that leads to the formation of cysteine, glutathione, and other metabolically important metabolites.[38] The rat cystathionine beta-synthase gene uses alternative exons to form four distinct mRNA isoforms, all sharing the middle portion but differ in the amino- and carboxyl-terminal sequences.[39] The human cystathionine beta-synthase gene uses multiple transcription initiation sites to yield at least five mRNA isoforms differing at their 5′ ends.[40] The original protein of *A. flavus* cystathionine beta-synthase was composed of four exons but the AS protein had only two exons with alternative 5′ sites (Supplementary Figure 1). Peptide TRTSDLPSTLQPHEQK (spectrum see Supplementary Figure 1) confirmed the boundaries and linkup of the two exons for the AS protein, where peptides AIVAGAGTGGTITGLSR and DYNFGKDDVVVVILPDSIR showed the original protein used different splice sites and translation activity occurred in the intron of the AS protein. The transcription start sites used by the two proteins were different and not in-frame. The different reading frame explained that peptide MQLSSALLMK was only found in the AS protein. The two reading frames became in-frame later in the translation, as evidenced by the shared peptide ISEVVTDPR.

Pyruvate decarboxylase (EC 4.1.1.1) catalyzes the decarboxylation of pyruvic acid to acetaldehyde and carbon dioxide. Pyruvate decarboxylase has been reported in obligate aerobic

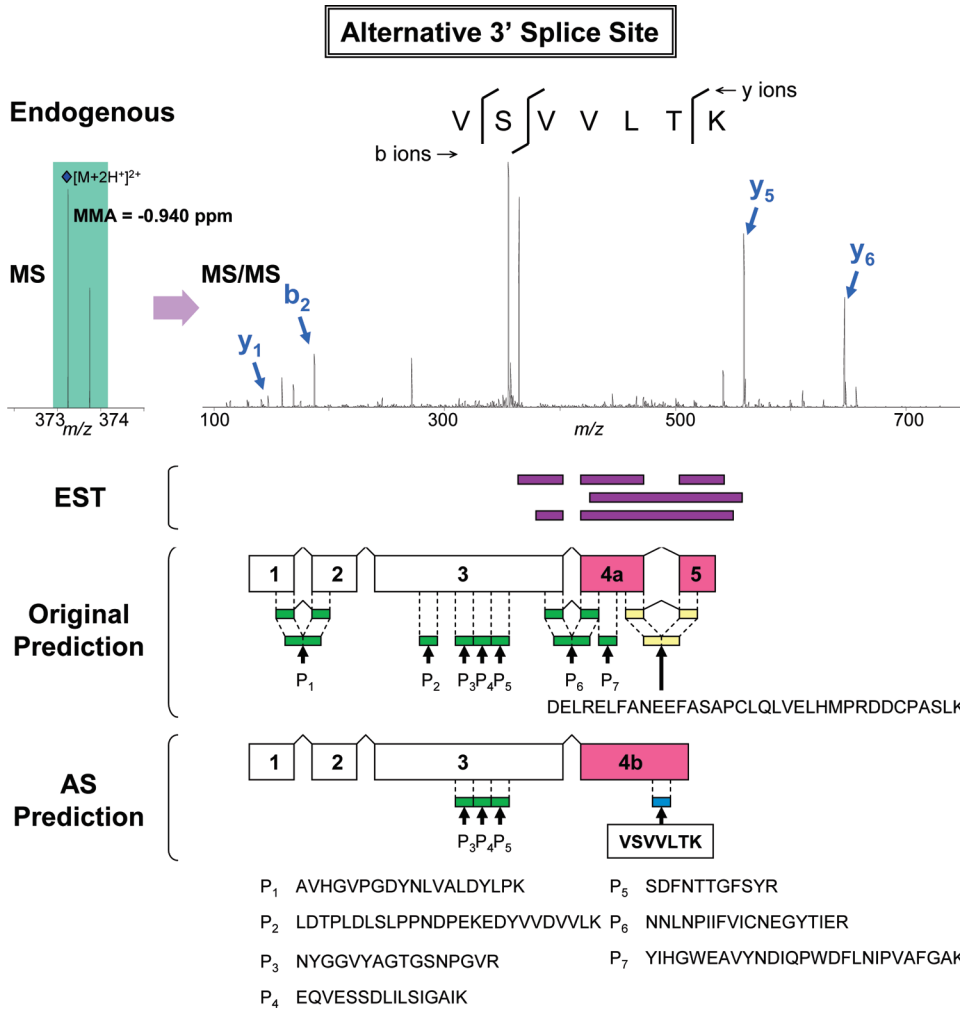

Figure 4. Pyruvate decarboxylase PdcA. The alternative 3′ splice site of the last exon in the AS protein also served as an alternative stop site (lower trace). The MS/MS spectrum of the AS protein-specific peptide VSVVLTK (upper trace) was resulted from a precursor ion scan at *m/z* 373.24 with a measured mass accuracy of 0.94 ppm. The AS protein-specific peptide located in the intron of the original protein suggested the alternative 3′ splice site. Original protein-specific, AS protein-specific, and indistinguishable shared peptide were labeled in yellow, blue and green, respectively.

filamentous fungi like *Aspergillus nidulans*.[41] Three pyruvate decarboxylase isozymes, encoded by three structural genes PDC1, PDC5 and PDC6, have been found in yeast.[42] The first three exons were constitutive exons in the original and AS proteins of *A. flavus* pyruvate decarboxylase PdcA, and the 3′ splice sites of the fourth exon were alternative (Figure 4). For the original protein, peptide DELRELFANEEFASAPCLQLVEL-HMPRDDCPASLK was derived from the splicing and joining of the fourth and fifth exons, indicating one alternative 3′ splice sites of exon four. For the AS protein, the coding region for the peptide VSVVLTK (spectrum see Figure 4) appeared to be inside the intron of the original protein, indicating an elongated exon four with another alternative 3′ site as well as an alternative stop site.

ATP synthase/ATPase (EC 3.6.3.14) is a ubiquitous enzyme consisting of two components, an extrinsic globular domain called F1 and a membrane intrinsic domain known as F0, linked together by a central and a peripheral stalk.[43] F1 is the catalytic domain made of subunits. The mitochondrial F1 ATPase subunit alpha gene is conserved in Eukaryota including human, mouse, rat, chicken, fruit fly, worm, yeast, and plant. Alternatively spliced transcript variants of human ATP5A1 gene encoding the same protein have been identified. Besides the four constitutive exons, the last exon was intact in the original protein of *A. flavus* mitochondrial F1 ATPase subunit alpha but broken into two smaller exons in the corresponding AS protein (Supplementary Figure 2). The detection of the peptide EVAAFAQFGSDLDAATK suggested the intron in the AS protein was kept in the original protein. In addition, peptide RPRLPSR (spectrum see Supplementary Figure 2) specific to the AS protein and peptides EGQVSKETEASLKEIIQSFNK/EIIQSFNK specific to the original protein were mapped back to the same coding region. These unique peptides resulted from the change of reading frames which was caused by the intron retention in the original protein.

For 17 new proteins identified from the AS database, there was no detection of the corresponding protein from the original database. In that case, MS/MS peptides alone might not be sufficient to tell the finding was the product of alternative splicing or the revision of gene annotation.

Discussion

The goal of this study was to test the following hypotheses: (i) whether tandem-MS spectra with high precursor ion mass accuracy is able to detect the existence of alternative variants

in *A. flavus*; and (ii) whether a computationally constructed AS database can explain more proteomic data by providing novel and putative protein isoforms for database searches. Although Mascot supports the use of nucleotide databases besides protein databases, the searching of the EST database has longer search time and higher false positive rate.[1] We constructed a tailor-made AS database of the filamentous fungus *A. flavus* based on public EST sequences. These computationally predicted sequences legitimately increased the target space of MS/MS peptide search by 76%. Besides 556 proteins identified from the original database, searching the expanded protein database was able to identify 29 new proteins encoded by 26 genes from the same data. The increase in the number of proteins identified from original to that of the AS database supported the hypothesis regarding to practical value of the AS database. Traditionally, the confirmation of alternatively spliced isoforms relies on the detection of different mRNA transcripts. By taking advantage of existing tandem MS data and an appropriate AS database, alternatively spliced variants can be detected more efficiently at the translation level in a large scale.

Considering a sample size of 556 original proteins identified by tandem MS analysis instead of the total 12 832 annotated genes of *A. flavus*, identifications of multiple protein variants of nine genes suggested that 1.6% of fungal genes were estimated to be alternatively spliced. The number is far less than the estimates in mammals like 40−60% for human genome but close to the scale of the estimated 4.2% in the basidiomycetous yeast.[9] The observation supported the hypothesis regarding the choice of the fungus model and the sufficient sensibility of the tandem MS experiment.

For 65% (17 out of 26) of the *A. flavus* genes having new proteins identified from the AS database, the corresponding original proteins were not found in the search results. Without a comparison basis to ascertain whether alternative splicing occurred or not, multiple plausible theories may explain the results equally well. One possibility is the original and AS proteins are both expressed in *A. flavus*. In the first plausible scenario, the protein identified from the AS database is genuinely a different isoform. No MS/MS spectrum is recorded to infer the existence of the corresponding original protein. Another plausible explanation is the previously predicted gene model is flawed or incomplete. The identification of the AS protein turns out to be the revision of previous annotation errors.

Current genome annotations are commonly generated by a computational pipeline and could contain errors. Recently, several research groups have already taken advantage of shotgun proteomic data to improve annotated genomes, including *Homo sapiens*,[44] *Arabidopsis thaliana*,[45] *Drosophila melanogaster*,[46] and *Caenorhabditis elegans*.[47] It is consistently reported that many peptides were mapped to genome sequences not considered as transcription regions previously. In fact, the most recent annotation of *A. flavus* genome (May 27, 2009) showed that three gene models in Table 1 had been updated and four peptides discovered in this study can be found in the latest gene models. This suggested that the peptides identified from the AS database are correct and significant. Although we cannot definitively rule out the possibility that some identifications as splice variants might be annotation errors, even with all the efforts of validation, alternative splicing is the most plausible explanation for those genes which have two different proteins identified.

Without including putative protein isoforms, the conclusions from proteomic studies in higher eukaryotic organisms might be providing only part of the biological picture. An AS database which can be easily integrated into MS-based analysis pipelines will fill this void and provide new biological insights. Although the gene models of some identified proteins remain to be clarified, our bioinformatic efforts help lift the value of existing experimental spectra and attain a better understanding of the protein profile in the organism of interest. This study demonstrates that alternative splicing should be taken into consideration for the analysis and interpretation of proteomic data.

Acknowledgment. The authors thank the W.M. Keck Foundation and North Carolina State University for supporting this research.

Supporting Information Available: Supplementary Figure 1, cystathionine beta-synthase; Supplementary Figure 2, mitochondrial F1 ATPase subunit alpha. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D. Informatics for protein identification by mass spectrometry. *Methods* **2005**, *35* (3), 223–236.
(2) Roth, M. J.; Forbes, A. J.; Boyne, M. T., II; Kim, Y. B.; Robinson, D. E.; Kelleher, N. L. Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (7), 1002–1008.
(3) Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **2003**, *72*, 291–336.
(4) Modrek, B.; Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **2002**, *30* (1), 13–19.
(5) Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **2006**, *126* (1), 37–47.
(6) Pajares, M. J.; Ezponda, T.; Catena, R.; Calvo, A.; Pio, R.; Montuenga, L. M. Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol.* **2007**, *8* (4), 349–357.
(7) Johnson, J. M.; Castle, J.; Garrett-Engele, P.; Kan, Z.; Loerch, P. M.; Armour, C. D.; Santos, R.; Schadt, E. E.; Stoughton, R.; Shoemaker, D. D. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **2003**, *302* (5653), 2141–2144.
(8) Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **2000**, *103* (3), 367–370.
(9) Galagan, J. E.; Henn, M. R.; Ma, L. J.; Cuomo, C. A.; Birren, B. Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.* **2005**, *15* (12), 1620–1631.
(10) Bairoch, A.; Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **1991**, *19*, 2247–2249, Suppl.
(11) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2007**, *35* (Database issue), D61–D65.
(12) Fermin, D.; Allen, B. B.; Blackwell, T. W.; Menon, R.; Adamski, M.; Xu, Y.; Ulintz, P.; Omenn, G. S.; States, D. J. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **2006**, *7* (4), R35.
(13) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–4399.
(14) Tabb, D. L.; Saraf, A.; Yates, J. R., III. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **2003**, *75* (23), 6415–6421.
(15) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–4639.
(16) Leipzig, J.; Pevzner, P.; Heber, S. The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res.* **2004**, *32* (13), 3977–3983.
(17) Holste, D.; Huo, G.; Tung, V.; Burge, C. B. HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.* **2006**, *34* (Database issue), D56–D62.

(18) Kim, P.; Kim, N.; Lee, Y.; Kim, B.; Shin, Y.; Lee, S. ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.* **2005**, *33* (Database issue), D75–D79.
(19) Kim, N.; Alekseyenko, A. V.; Roy, M.; Lee, C. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.* **2007**, *35* (Database issue), D93–D98.
(20) Koscielny, G.; Le Texier, V.; Gopalakrishnan, C.; Kumanduri, V.; Riethoven, J. J.; Nardone, F.; Stanley, E.; Fallsehr, C.; Hofmann, O.; Kull, M.; Harrington, E.; Boue, S.; Eyras, E.; Plass, M.; Lopez, F.; Ritchie, W.; Moucadel, V.; Ara, T.; Pospisil, H.; Herrmann, A.; J, G. R.; Guigo, R.; Bork, P.; Doeberitz, M. K.; Vilo, J.; Hide, W.; Apweiler, R.; Thanaraj, T. A.; Gautheret, D. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* **2009**, *93* (3), 213–220.
(21) Shah, P. K.; Jensen, L. J.; Boue, S.; Bork, P. Extraction of transcript diversity from scientific literature. *PLoS Comput. Biol.* **2005**, *1* (1), e10.
(22) Yu, J.; Cleveland, T. E.; Nierman, W. C.; Bennett, J. W. Aspergillus flavus genomics: gateway to human and animal health, food safety, and crop resistance to diseases. *Rev. Iberoam. Micol.* **2005**, *22* (4), 194–202.
(23) Squire, R. A. Ranking animal carcinogens: a proposed regulatory approach. *Science* **1981**, *214* (4523), 877–880.
(24) Wogan, G. N. Aflatoxins as risk factors for hepatocellular carcinoma in humans. *Cancer Res.* **1992**, *52*, 2114s–2118s, 7 Suppl.
(25) Hedayati, M. T.; Pasqualotto, A. C.; Warn, P. A.; Bowyer, P.; Denning, D. W. Aspergillus flavus: human pathogen, allergen and mycotoxin producer. *Microbiology* **2007**, *153* (Pt. 6), 1677–1692.
(26) Georgianna, D. R.; Hawkridge, A. M.; Muddiman, D. C.; Payne, G. A. Temperature-dependent regulation of proteins in *Aspergillus flavus*: whole organism stable isotope labeling by amino acids. *J. Proteome Res.* **2008**, *7* (7), 2973–2979.
(27) Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. A.; Muddiman, D. C. Top-down identification and quantification of stable isotope labeled proteins from *Aspergillus flavus* using online nano-flow reversed-phase liquid chromatography coupled to a LTQ-FTICR mass spectrometer. *Anal. Chem.* **2008**, *80* (13), 4994–5001.
(28) Kubodera, T.; Watanabe, M.; Yoshiuchi, K.; Yamashita, N.; Nishimura, A.; Nakai, S.; Gomi, K.; Hanamoto, H. Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Lett.* **2003**, *555* (3), 516–520.
(29) Payne, G. A.; Nierman, W. C.; Wortman, J. R.; Pritchard, B. L.; Brown, D.; Dean, R. A.; Bhatnagar, D.; Cleveland, T. E.; Machida, M.; Yu, J. Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Med. Mycol.* **2006**, *44*, 9–11, Suppl.
(30) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.
(31) Florea, L.; Hartzell, G.; Zhang, Z.; Rubin, G. M.; Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **1998**, *8* (9), 967–974.
(32) Heber, S.; Alekseyev, M.; Sze, S. H.; Tang, H.; Pevzner, P. A. Splicing graphs and EST assembly problem. *Bioinformatics* **2002**, *18* (Suppl. 1), S181–S188.
(33) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **1993**, *3* (6), 327–332.
(34) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res.* **2008**, *7* (1), 29–34.
(35) Johnson, K. L.; Ovsyannikova, I. G.; Poland, G. A.; Muddiman, D. C. Identification of class II HLA-DRB1*03-bound measles virus peptides by 2D-liquid chromatography tandem mass spectrometry. *J. Proteome Res.* **2005**, *4* (6), 2243–2249.
(36) Degtyarenko, K. N.; Kulikova, T. A. Evolution of bioinorganic motifs in P450-containing systems. *Biochem. Soc. Trans.* **2001**, *29* (Pt. 2), 139–147.
(37) Borgese, N.; D'Arrigo, A.; De Silvestris, M.; Pietrini, G. NADH-cytochrome b5 reductase and cytochrome b5 isoforms as models for the study of post-translational targeting to the endoplasmic reticulum. *FEBS Lett.* **1993**, *325* (1−2), 70–75.
(38) Miles, E. W.; Kraus, J. P. Cystathionine beta-synthase: structure, function, regulation, and location of homocystinuria-causing mutations. *J. Biol. Chem.* **2004**, *279* (29), 29871–29874.
(39) Swaroop, M.; Bradley, K.; Ohura, T.; Tahara, T.; Roper, M. D.; Rosenberg, L. E.; Kraus, J. P. Rat cystathionine beta-synthase. Gene organization and alternative splicing. *J. Biol. Chem.* **1992**, *267* (16), 11455–11461.
(40) Kraus, J. P.; Oliveriusova, J.; Sokolova, J.; Kraus, E.; Vlcek, C.; de Franchis, R.; Maclean, K. N.; Bao, L.; Bukovsk; Patterson, D.; Paces, V.; Ansorge, W.; Kozich, V. The human cystathionine beta-synthase (CBS) gene: complete sequence, alternative splicing, and polymorphisms. *Genomics* **1998**, *52* (3), 312–324.
(41) Lockington, R. A.; Borlace, G. N.; Kelly, J. M. Pyruvate decarboxylase and anaerobic survival in *Aspergillus nidulans*. *Gene* **1997**, *191* (1), 61–67.
(42) Pronk, J. T.; Yde Steensma, H.; Van Dijken, J. P. Pyruvate metabolism in *Saccharomyces cerevisiae*. *Yeast* **1996**, *12* (16), 1607–1633.
(43) Walker, J. E.; Dickson, V. K. The peripheral stalk of the mitochondrial ATP synthase. *Biochim. Biophys. Acta* **2006**, *1757* (5−6), 286–296.
(44) Tanner, S.; Shen, Z.; Ng, J.; Florea, L.; Guigo, R.; Briggs, S. P.; Bafna, V. Improving gene annotation using peptide mass spectrometry. *Genome Res.* **2007**, *17* (2), 231–239.
(45) Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **2008**, *320* (5878), 938–941.
(46) Brunner, E.; Ahrens, C. H.; Mohanty, S.; Baetschmann, H.; Loevenich, S.; Potthast, F.; Deutsch, E. W.; Panse, C.; de Lichtenberg, U.; Rinner, O.; Lee, H.; Pedrioli, P. G.; Malmstrom, J.; Koehler, K.; Schrimpf, S.; Krijgsveld, J.; Kregenow, F.; Heck, A. J.; Hafen, E.; Schlapbach, R.; Aebersold, R. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat. Biotechnol.* **2007**, *25* (5), 576–583.
(47) Merrihew, G. E.; Davis, C.; Ewing, B.; Williams, G.; Käll, L.; Frewen, B. E.; Noble, W. S.; Green, P.; Thomas, J. H.; MacCoss, M. J. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Res.* **2008**, *18* (10), 1660–1669.

PR900602D