

Virtual Polymorphism: Finding Divergent Peptide Matches in Mass Spectrometry Data

Rebekah Starkweather,[†] Charles S. Barnes,[‡] Gerald J. Wyckoff,[†] and J. Andrew Keightley^{*,†}

Division of Molecular Biology and Biochemistry, University of Missouri—Kansas City, 5007 Rockhill Road, Kansas City, Missouri 64110, and Allergy/Asthma/Immunology Research Lab, Children's Mercy Hospitals & Clinics, 2401 Gillham Road, Kansas City, Missouri 64108

The prevailing method of analyzing tandem-MS data for protein identification involves the comparison of peptide molecular weight and fragmentation data to theoretically predicted values, based on known protein sequences in databases. This is generally effective since proteins from most species under study are in the database or have sufficient homology to allow significant matching. We have encountered difficulties identifying proteins from fungal species *Alternaria alternata* due to significant inter-species protein sequence differences (divergence) and its absence from the database. This common household mold causes asthma and allergy problems, but the genome has not been sequenced. De novo sequencing and error-tolerant methods can facilitate protein identifications in divergent, unsequenced species. But these standard methods can be laborious and only allow single amino acid substitution, respectively. We have developed an alternative approach focusing on database engineering, predicting biologically rational polymorphism using statistically weighted amino acid substitution information held in BLOSUM62. Like other second pass methods, it is based on the initially identified protein. However, this approach allows more control over sequences to be considered, including multiple changes per peptide. The results show considerable improvement for routine protein identification and the potential for rescuing otherwise unconvincing identifications in unusually divergent species.

The introduction of automated protein identification based on the comparison of peptide mass spectral data to peptide mass and fragmentation patterns calculated from existing database entries revolutionized protein identification and provided impetus for the rapid development of the field of proteomics.¹ Many database search applications² are now available commercially or through open source code for both MS only and tandem-MS data.^{3–8}

Tandem-MS approaches (MS/MS) have the advantage of allowing confident protein identification on the basis of only a few peptide spectra, or even one match, depending on the quality of the data.⁹ One of the most powerful attributes of the database matching strategy is that even though a spectrum may be missing fragment peaks and have significant detector noise, a score based on conforming precursor mass, the assignment of abundant peaks, and a discernible b- or y-series is usually sufficient to allow that match to rise well above random matches using any one of a variety of searching algorithms and scoring methods available.² This makes the database matching strategy intrinsically effective and even permits substantial error in mass assignments routinely produced by ion trap and tandem quadrupole instrumentation.

Commonly studied species are represented in the sequence databases, and minor sequence differences are usually compensated for by the identification of conserved peptides in the existing entries. However, we are investigating the proteome of *Alternaria alternata*, a common mold that is poorly represented in the protein database. *Alternaria* are a phenotypically diverse group of fungi first described in the 1800s; mycologists do not know the precise number of *alternaria* species, but several hundred have been described. They are ubiquitous and are an allergen for humans and an emerging pathogen.^{10–12} *A. alternata* is also a noted plant pathogen, causing leaf, fruit, and other rots on over 380 hosts, including many crop species. Currently, there is no large-scale sequencing project of *Alternaria* species; within *A. alternata*, there are approximately 63 unique proteins represented within NCBI's protein database. Sequence identity with other species that are well-represented within the database (for example, *Aspergillus* and *Neurospora*) range from almost 99% for well-conserved proteins (such as histones) to less than 50% for many enzymes (from a survey of these 63 proteins; data not shown). Distribution of amino

* To whom correspondence should be addressed. Phone: (816) 235-2565. E-mail: keightleyj@umkc.edu.

[†] University of Missouri—Kansas City.

[‡] Children's Mercy Hospitals & Clinics.

(1) Yates, J. R., III.; et al. *Anal. Biochem.* **1993**, *214*, 397–408.

(2) Sadygov, R. G.; Cociorva, D.; Yates, J. R., III. *Nat. Methods* **2004**, *1*, 195–202.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass. Spectrom.* **1994**, *5*, 976–989.

(4) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cotrell, J. C. *Electrophoresis* **1999**, *20*, 3551–3567.

(5) Field, H. I.; Fenyö, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36–47.

(6) Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406–1412.

(7) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466–1470.

(8) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958–964.

(9) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, C.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531–533.

(10) Barnes, C.; Portnoy, J.; Sever, M.; Arbes, S.; Vaughn, B.; Zeldin, D. *Ann. Allergy Asthma Immunol.* **2006**, *97*, 1–7.

(11) Curran, I.; Young, N.; Burton, M.; Vijay, H. *Int. Arch. Allergy Immunol.* **1993**, *102*, 267–275.

(12) Paris, S.; Debeaupuis, J. P.; Prevost, M. C.; Casotto, M.; Latge, J. P. *J. Allergy Clin. Immunol.* **1991**, *88*, 902–908.

acid sequence variation throughout *A. alternata* prevents a substantial number of peptide matches that might otherwise be found in the homologues that are present in the database from other species. This reduces the number of peptide matches with which to base a protein identification. *A. alternata* is a typical example of a species that cannot take full advantage of the standard database matching search strategy.

To find the additional peptide matches in the data and enhance confidence in a tentative protein identification, there are "second pass" search strategies available, typically focusing on the putatively identified protein sequence. Various methods of error-tolerant searching (ETS) interrogate the data while allowing for amino acid substitution (discussed below) or amino acid mass adjustments for selected post-translational modification. Another strategy allows for protease nonspecificity, unexpected source fragmentation, or biological protein processing by permitting no enzyme specificity (both ends) or semienzyme specificity (one end) in the peptides considered from the tentatively identified protein sequence. Alternatively, manual or automated de novo spectral interpretation can be attempted. This process involves solving spectra through computational means or at least building a partial sequence (sequence tag) based on a discernible series and looking for the sequence in the tentatively identified protein. Coupling this strategy with filling in the incomplete peptide, allowing single amino acid substitution or modification to bridge the gap to the measured precursor mass can also generate potential matches.^{13,14} There are additional strategies that use interpreted sequence in various ways for protein identification, but automated de novo applications vary in their ability to accurately solve spectra.¹⁵ Complete de novo interpretation can be successful, particularly for shorter peptides. As a peptide grows in length, however, the number of mathematically compatible sequences also grows. Compounding this, absent peaks, poor mass accuracy, and spectral noise presents misleading data to the algorithms or to the human interpreter. Peptide excitation strategies that allow more uniform fragmentation are now more readily available to eliminate missing peaks (electron capture dissociation, for example), and accurate mass measurement technology is also becoming more common. However, the majority of proteomics facilities still rely on unmodified ion traps, tandem quadrupole, or Q-tof mass spectrometers for the majority of acquired proteomics data.

When minor sequence differences are present, ETS strategies that allow for amino acid substitution can be implemented. Traditionally, most have relied on single-nucleotide polymorphism (SNP) to predict substitutions. The SNP table presents amino acid substitutions based on a single change in any position of any codon to all three possible alternative nucleotides and has been commonly used for error-tolerant approaches.^{13,16,17} The table is about 2.5 times smaller than a comprehensive amino acid substitution

table (20 × 20), but it is not statistically weighted (substitution is allowed, or it is not allowed), and it does not foresee the effect of subsequent mutation(s), which would be relevant in more divergent species. DNA mutation is based on DNA polymerase infidelity or damage by a variety of events, with repair mechanisms that have inherently unique, in some cases, sequence-specific error rates.^{18,19} Therefore, some mutations are substantially more likely to occur than others, and the resulting changes are then subjected to natural selection. The reality of these more complex patterns is not reflected in the SNP table, but is readily available in other matrices that score the actual frequency of substitution events observed in nature. The work presented here represents our first step in designing and using a flexible second pass search method that takes advantage of the inherent strengths of the database matching strategy and the benefits of amino acid substitution prediction based on observed rates.

BLOSUM was originally designed and used for scoring the similarity between more distantly related species²⁰ and is frequently used to find related DNA or protein sequences, such as in BLAST searches. The values in a BLOSUM62 matrix are based on observed substitution events seen in alignments between sequences having no more than 62% identity. Therefore, the statistically weighted amino acid substitution information held in the BLOSUM62 matrix should also be able to predict protein sequences that may be present in the more divergent species. In the current work, BLOSUM62 was used to systematically generate defined single and double substitutions in searchable peptides, based on the sequence of a tentatively identified protein, with the resulting peptides compiled as separate entries into a database for second pass searching. Fifty databases were generated for 50 *A. alternata* proteins that had been tentatively identified from existing database entries from other fungal species. The databases were interrogated for new matches, resulting in substantial improvement of most of the protein identifications by more than doubling, on average, the number of identified peptides. Many of these newly identified peptides have two substitutions with respect to the original reference sequence, making them difficult or impossible to identify by any other available method. This strategy offers an alternative method for improving protein identifications in an unsequenced species.

EXPERIMENTAL SECTION

Sample Preparation, Mass Spectrometry, and Initial Protein Identification. All reagents were from Sigma (St. Louis, MO), unless otherwise noted. Isoelectric focusing, SDS-PAGE supplies, and DC protein assay reagents were from Bio-Rad (Hercules, CA). *A. alternata* (ATCC 96154) was grown for 2 weeks on chemically defined medium [0.3% NaNO₂, 0.05% MgSO₄, 0.05% KCl, 0.1% K₂HPO₄, 3% sucrose, and trace elements consisting of 1.8 mM Fe(NO₃)₃, 0.97 mM ZnSO₄, and 1.8 mM MnSO₄].^{21,22} Total

- (13) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390–4399.
- (14) Beavis, R. C. Using the Global Proteome Machine for Protein Identification, In *New and Emerging Proteomic Techniques*; Nedelkov, D., Nelson, R. W., Eds.; Methods in Molecular Biology 328; Humana Press: Totowa, NJ, 2006; pp 217–228.
- (15) Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X. *J. Proteome Res.* **2006**, *5*, 3018–3028.
- (16) Roth, M. J.; Forbes, A. J.; Boyne, M. T., II; Kim, Y.-B.; Robinson, D. E.; Kelleher, N. L. *Mol. Cell. Proteomics* **2005**, *4*, 1002–1008.
- (17) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R., III. *Anal. Chem.* **2000**, *72*, 757–763.

- (18) *DNA Repair and Mutagenesis*; Friedberg, E. C., Walker, G. C., Siede, W., Eds.; ASM Press: Washington DC, 1995.
- (19) *DNA Damage and Repair*, Volume I: *DNA Repair in Prokaryotes and Lower Eukaryotes*, and Volume II: *DNA Repair in Higher Eukaryotes*; Nickoloff, J. A., Hoekstra, M. F., Eds.; Humana Press: Totowa, NJ, 1998.
- (20) Henikoff, S.; Henikoff, J. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915–10919.
- (21) Schaffer, N.; Molomut, N.; Center, J. G. *Ann. Allergy* **1959**, *380*, 17.
- (22) Bush, R.; Recht, M.; Kroutil, L.; Sweeney, M. A. *J. Allergy Clin. Immunol.* **1985**, *52* (abstr), 75.

mycelial protein from *A. alternata* mycelial culture was homogenized (Polytron, Brinkman Instruments, Westbury, NY) with three 45-s bursts in lysis buffer containing 20 mM TrisCl, 50 mM NaCl, 2 mM MgCl₂, pH 7.5, 0.5% TX-100, 20 µg/mL pancreatic RNAase A, 20 µg/mL bovine DNAase I, and protease inhibitor (Roche Complete mini tablet). After centrifugation (20000g, 30 min), protein yield was estimated in supernatants (DC protein assay). 2D gels were run as described.²³

Picofrit columns (50 µM i.d., New Objective) packed with Phenomenex Jupiter C18 matrix (7–8 cm in length), and an Eksigent binary capillary liquid chromatography system were used to generate a 45-min reversed-phase gradient (2–70% acetonitrile, in 0.1% acetic acid, 200 nL/min) to resolve peptides for analysis. LC tandem-MS data was collected on a Thermo Finnigan LTQ as described²³ but with eight dependent scans per cycle rather than three. Resulting data files were searched with both Sequest (Thermo) and Mascot (Matrix Science, Ltd.) for the sole purpose of obtaining preliminary identifications based on known fungal proteins in the existing databases. Regardless of the originating species for the preliminary match, the highest scoring candidate (the reference protein sequence) from an existing database entry was used as input to generate a polymorphism-containing database as described below. This choice was made on the assumption that, in general, the best scoring fungal species match is likely to be closely related to the corresponding *Alternaria* sequence and is therefore likely to be a productive starting sequence for this strategy. Subsequently, all reference protein sequences and virtual polymorphism-containing databases were searched using only Sequest as described below for the performance evaluation.

Database Engineering. The BLOSUM62 matrix (http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/BLOSUM62) was used to allow amino acid substitution based on the value in the matrix for each amino acid pair. Values range from −4 (very uncommon substitution) to 3 (very common substitution). Databases were generated using a program operating through a web-based interface that accepts a reference amino acid sequence as input (MS-Virtmorph, available at <http://ms-virtmorph.umkc.edu/>). The program requires the user to select depth of allowed substitution (from −4 to 3), to select the protease to be used in the subsequent search (required for generating output peptides), to select optional iterative substitution (up to three), and offers user-defined position exclusion (known conserved positions, for instance). The option to include an amino acid extension to the next protease site and to the previous site with each peptide entry in the database (using adjacent reference protein amino acid sequence; see example below) is also allowed and was used in this work. This option makes it possible to use the missed cleavage option while searching an engineered database.

For each reference protein, a database of virtual peptides containing two rounds of substitution at allowable depth of −1 or above was engineered, excluding only the first position (methionine), retaining the intermediate singly substituted entries in each database generated. Thus, when glycine is encountered, substitution at that position with alanine (0), asparagine (0), aspartate (−1), or serine (0) is allowed (but not arginine (−2), for instance).

Met to Phe was not allowed, but searches included oxidation of methionine (Met^{ox}) to accommodate this potential mass change (+16 Da). To minimize the potential effect of a combinatorial explosion, Met/Met^{ox} was the only differential modification included in searches. Evidence for neutral loss (CH₃-SOH) from Met^{ox} peptide precursor ion or fragment ions was used to distinguish between presence of oxidized Met^{ox} or Phe during validation. Single dalton changes based on N = D and Q = E were also allowed (which inadvertently accommodates deamidation events, if present: N > D or Q > E). E = K is another example of a 1-Da change that was allowed, but with the added significance of potentially effecting trypsin specificity. Q = K substitutions were also allowed, though near isobaric, due to introduction/removal of a standard tryptic site.

Building Virtual Peptide Entries. When executed, the program systematically generates a separate peptide entry for each allowed substitution in the first round, working with each canonical peptide based on protease specificity selected (trypsin, in this work). Progressing through the sequence from N- to C-terminus, a comprehensive database with separate peptide entries containing single substitutions is produced in the first round, optional extensions to the next protease (trypsin) site in both directions are added (from reference sequence; see example below), and a fasta definition line is generated to precede each entry. The definition line describes the entry, including matrix and depth of allowed substitution selected, amino acid position based on the reference sequence numbering, and the imposed substitution. In the subsequent round(s), the program takes each substituted peptide entry (without extensions), holds the substitution constant, and generates a series of entries with second substitutions, with the optional reference sequence extensions reinstated to build each entry. The definition lines are modified to reflect the new substitutions. The full-length reference sequence is retained as the first entry in the database. Here is an example entry taken from ketol-acid reductoisomerase:

```
>NcRedIsom.102010|1B62-1|967|223|F=L|2B62-1|8649|226|Y=F|
EGRGINSSLAVFQDVTGKAK
```

The above entry is built around the tryptic peptide GINSS-FAVYQDVTGK from the reference species *Neurospora crassa*, with extensions to the next trypsin site in both directions (identical to the corresponding region of the reference sequence). The definition line shows this is entry number 102010 in this database, with two substitutions described at positions 223 (Phe to Leu) and 226 (Tyr to Phe). Also visible is the first and second substitution criteria chosen for database engineering (1B62-1 and 2B62-1, respectively). The other numbers are counters used by the program. Since there are no spaces in the definition line, Sequest (BioWorks 3.1 SR1, Thermo Electron) retains it as the name of the entry, showing the change(s) that allow a match to occur in search results.

Virtual Polymorphism-Containing Database and Reversed Database Searches. Completed MS-Virtmorph databases (VM databases) were saved (.fasta file extension), and the same LC tandem-MS data file that generated the initial match was reintegrated for new matches using Sequest (BioWorks 3.1 SR1). Searches allowed tolerances (average mass) of 1.8 Da for peptide mass, 1 Da for fragment mass, 1 missed cleavage, fixed modifica-

(23) Keightley, J. A.; Shang, L.; Kinter, M. *Mol. Cell. Proteomics* **2004**, *3*, 167–175.

tion of cysteine (carbamidomethylation), and differential modification of methionine (oxidation). Search times varied from less than 1 h to 90 min on a Pentium IV, 3GHz, with 2 Gb main memory. An additional separate search was conducted on the reference sequence alone using identical search parameters to verify the maximum number of peptide matches available in the reference sequence. Matches based on an assumption of +1 charge state were not considered or scored in any of the analysis, due to the relative low quality of such spectra in ion trap data, and the increased potential for random matches that shorter peptides inherently possess.

Results from all VM database searches were initially filtered to meet a threshold Xcorr score of 2.400 for +2 matches, and 3.500 for +3 matches. The resulting list was exported from Bioworks as xml and parsed using a perl script (available upon request). The parser organized the results with best single scoring virtual sequence match for each scan presented first (unless matched to a peptide in the reference sequence), matches assigned to the reference sequence next, and all matches that met basic threshold criteria next (both reference and virtual sequences, by scan, ascending). The Bioworks search summary and the parsed search results were used during the validation.

Reversed sequence databases were produced by inverting the order of each peptide entry in a given database and recompiling the resulting entries into a database with the letter R added to each definition line. Reversed databases were saved with and without the forward database appended, and both versions were searched. The same search parameters and Xcorr threshold criteria described above were applied to limit the output of the searches. For the reversed database searches, no validation or exclusion was implemented following the search besides basic threshold criteria (Xcorr versus charge state, as above).

For scoring purposes, new matches in the forward VM databases, however, were summarily rejected for a variety of reasons, including the following: (a) if the scan matched a peptide in the reference sequence, (b) if any part of the matched virtual sequence overlapped a sequence represented in the matches to the reference (including partial digest, etc.), and (c) if any apparent conflicting information raised a question about the match, such as unresolved oxidation compliance (± 6 Da) or unlikely trypsin activity (multiple acidic residues at trypsin site). Manual inspection of the remaining matches considered standard validation procedures,^{24,25} such as the presence of proline, histidine, or acidic residue that might result in an expected predominant fragment(s), expected charge state of precursor, anticipated neutral loss from precursor or daughter ion from serine/threonine (H_2O), oxidized methionine (CH_3-SOH), etc. For triply charged spectra, a transition to doubly charged fragments was generally expected after crossing a histidine or proline in a growing b- or y-series. All acceptable matches were retained with summary information for subsequent overall scoring performance evaluation of the searches.

RESULTS AND DISCUSSION

Our efforts were driven by a fundamental difficulty in assigning peptide matches and identifying proteins in the divergent, unsequenced fungal species *A. alternata*. Our hypothesis was that by imposing statistically weighted amino acid substitutions on the known proteins initially identified, we could predict the novel peptide sequences that are represented and identifiable in the LC tandem-MS data. Since amino acid substitution matrices are derived from the incidence of substitutions found in the comparison of existing related sequences^{26,27} the likelihood of success seemed quite high. These matrices have been employed successfully in local and global alignment programs^{28,29} and evolutionary studies.^{30,31} To facilitate database engineering and provide a template for developing this strategy, we built a web-based user interface, which allows control over a number of variables including depth of imposed substitution and optional iterative substitution. Including both singly and doubly substituted peptides, and allowing all substitutions above a score of -2 in the BLOSUM62 matrix, the databases engineered in this work should predict common substitutions as well as fairly divergent novel peptides. Output from this program is fasta formatted and should be portable to all major search programs used for protein identification.

Performance Evaluation: Protein Sequence Prediction Using BLOSUM62. Fifty databases were generated for 50 proteins that had been initially identified in 46 samples based on matches to existing database entries. The resolved proteins that obtained new peptide matches are indicated in the 2D gel shown in Figure 1, including three secondary identifications. The number of peptides matching the existing database entry and the number of new peptides found in searching virtual polymorphism databases is summarized for the 39 proteins listed in Table 1. The table shows the protein name, NCBI entry (gi), and species that seeded database engineering. The databases ranged in size from 43 108 (3.4 Mb) to 994 604 entries (130 Mb), with an average of 286 571 entries (5595 entries/kDa). There were 11 proteins that failed to obtain any acceptable new matches in the experiment (Table S-1, Supporting Information). Many of these failed experiments are from relatively small proteins with a single peptide match. In these cases, it is possible that the process failed to predict correct sequences to allow new matches. Alternatively, the initial identification may have been wrong, or no additional peptide MS/MS spectra were available for matching. A comprehensive list including the sequence of all 160 new peptides found in the data, their best score (Xcorr), the number of scans matched to each, substitutions that allowed the match, substitution depth required, the charge state(s) isolated for MS/MS, and additional information is available in Table S-1 and summarized in the following figures.

The successful prediction of naturally occurring sequences is summarized in Figure 2a, showing the average number of peptides per protein identified in NCBI database entries (existing) and the average number of new peptides per protein (new) discovered in

(24) Kinter, M.; Sherman, N. E. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*; Wiley-Interscience Series on Mass Spectrometry; John Wiley and Sons: New York, 2000.

(25) Chen, Y.; Kwon, S. W.; Kim, S. C.; Zhao, Y. J. *Proteome Res.* **2005**, *4*, 998–1005.

(26) Tang, H.; et al. *Mol. Biol. Evol.* **2004**, *21*, 1548–1556.

(27) Henikoff, S.; Henikoff, J. *Proteins Struct. Func. Genet.* **1993**, *17*, 49–61.

(28) Accelrys, The Wisconsin Package. 2003, Genetics Computer Group: Madison, WI, 2003.

(29) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403–410.

(30) Wyckoff, G. J.; Wang, W.; Wu, C. I. *Nature* **2000**, *403*, 304–309.

(31) Yang, Z.; Nielsen, R. *Mol. Biol. Evol.* **2000**, *17*, 32–43.

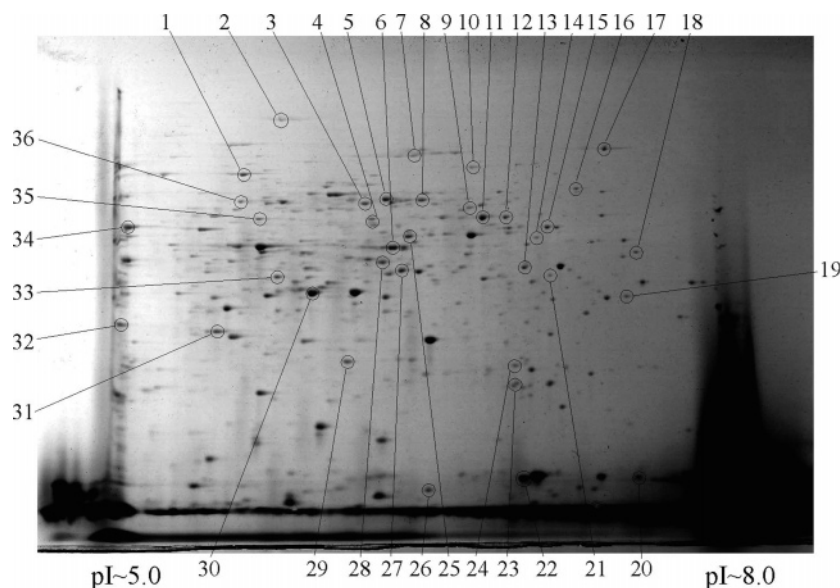


Figure 1. *A. alternata* total mycelial protein, separated by IEF and SDS-PAGE. Proteins that received new identified peptides based on virtual polymorphism are indicated. Corresponding protein identities are listed in Table 1.

the process. The existing database entries yielded 131 peptide matches in 39 proteins (3.359 peptides per protein), whereas an overall increase of 160 new peptides (4.103 per protein) was found in the virtual polymorphism-containing databases. For this divergent fungal species, the overall number of peptide matches found in the MS/MS data was more than doubled (from 3.359 to 7.462 peptides per protein). New matches by charge state (+2, +3) show that, even in this relatively small experiment, the new peptides were distributed similarly compared to the existing matches. Notably, peptides identified on the basis of matching MS/MS scans in both charge states (+2 and +3), which is possibly the most stringent criterion, showed similar relative gains.

New Peptide Matches Based on Single Substitution. Both singly and doubly substituted entries were present in the databases, and there are methods currently available to identify peptides with single amino acid changes. It is therefore relevant to distinguish matches made in these two categories. A total of 40 new peptides had single substitutions, whereas 120 doubly substituted peptides were identified (Table S-1). More substantial gains in the doubly substituted category are probably due to the divergent nature of this species, but note that whenever a single substitution afforded a reasonable match, it was preferentially selected if it endured manual inspection. In fact, singly substituted entries were selected and scored even if a doubly substituted peptide got a higher Xcorr score, because (1) if a single substitution allowed the match, double substitution was not necessary and (2) it is conceivable that a second substitution might result in a higher score by chance. Ion trap data has significant mass error, and a single dalton change (Asp to Asn, for instance) could enhance the score of a peptide inadvertently, or correctly, and it is unlikely to be easy to distinguish this routinely. It will be possible to differentiate these events more clearly when searching data acquired on instruments capable of higher mass accuracy, using narrower mass tolerances in the searches.

Of the 40 new peptide matches that were based on single substitution, 6 were not predicted by the SNP table that has been traditionally used for predicting polymorphisms. As mentioned

in the introduction, the SNP table predicts amino acid substitution based on all predicted nucleotide changes for every codon. Sequest-SNP,¹⁷ Mascot,³² Interrogator³³ (ProID), and ProSight-PTM¹⁶ accomplish this type of search, but not necessarily by the same mechanism. Sequest-SNP, for instance, imposes SNP changes in open reading frames and retranslates them to generate SNP-driven substituted peptides. Mascot ETS uses the SNP table to allow substitutions but also includes a full complement of modifications to be considered in the search. To our knowledge, all SNP implementations are limited to one substitution per peptide, which might be sufficient for most commonly studied species. But some predicted SNP mutations are actually quite rare. For instance, of the 75 forward substitutions predicted in the SNP table (and corresponding 75 backward substitutions), 12 receive a score of -3 in the BLOSUM62 matrix: $R = C$, $R = I$, $R = W$, $N = I$, $D = Y$, $D = V$, $C = G$, $G = V$, $H = L$, $I = K$, $L = P$, and $S = W$. These substitutions (scoring -3) can be 100 times less likely to occur than those receiving a -1 in the BLOSUM62 matrix (log odds score, rounded to integer).²⁰ Conversely, more than a dozen forward substitutions that are not allowed by the SNP approach (and their corresponding reverse substitutions) receive a score of 0 in the BLOSUM62 matrix and are thus more likely to occur than the SNP predicted substitutions shown above. These and other observations suggest that the use of event-based substitution matrices like BLOSUM to predict substitution and polymorphism should be more successful. The point mutation modeling capability offered in X-Tandem offers statistically weighted scoring for refinement of protein identifications based on accepted point mutations, or PAM.³⁴ When enabled, the point mutation modeling algorithm actually allows all amino acid substitution to take place, however, subsequently applying the PAM scoring filter

(32) Creasy, D. M.; Cotrell, J. S. *Proteomics* **2002**, *2*, 1426–1434.

(33) Tang, W. H.; Halpern, B. R.; Shilov, I. V.; Seymour, S. L.; Keating, S. P.; Loboda, A.; Patel, A. A.; Schaeffer, D. A.; Nuwaysir, L. M. *Anal. Chem.* **2005**, *77*, 3931–3946.

(34) Pearson, W. A. Rapid and Sensitive Sequence Comparison with FASTA and FASTP In *Methods in Enzymology*; Doolittle, R., Ed.; Academic Press: San Diego, CA, 1990; Vol. 183, pp 63–98.

Table 1. Summary of Proteins Identified^a

gel ID no.	protein	NCBI entry	species	size (kDa)	db entries	summary of existing and new peptides found	
						existing	new
34	ATP synthase β chain	GI:74613396	<i>N. crassa</i>	55.7	314167	12	9
17	methionine synthase	GI:74665265	<i>A. nidulans</i>	87.0	425718	10	4
6	adenosyl homocysteinase	GI:49651171	<i>Y. lipolytica</i>	49.7	358027	7	6
16	sulfate adenylyltransferase	GI:74618118	<i>N. crassa</i>	64.6	285356	6	9
23	GSP1/Ran	GI:54036030	<i>N. crassa</i>	24.4	132554	6	3
5	pyruvate kinase	GI:54036128	<i>N. crassa</i>	58.3	273742	5	9
1	HSP70/dnaK	GI:74629323	<i>P. brasiliensis</i>	73.9	386096	5	8
33	3-isopropylmalate DHase	GI:74625517	<i>P. nodorum</i>	39.0	264068	5	6
13	NAD-dependent formate Dhase	GI:4572458	<i>M. graminicola</i>	45.9	235218	5	5
32	14-3-3 homologue	GI:74662426	<i>N. crassa</i>	29.4	149535	5	2
27	ketol-acid reductoisomerase	GI:168822	<i>N. crassa</i>	44.7	219825	4	9
2	carbamoyl phosphate synthase	GI:49644360	<i>K. lactis</i>	123.7	994604	4	5
19	G protein β -like (WD repeat)	GI:52000514	<i>F. oxysporum</i>	35.4	208824	4	5
15	fumarate lyase	GI:74628637	<i>N. crassa</i>	57.5	370479	4	2
11	6-phospho-gluconate DHase	GI:28949952	<i>A. niger</i>	54.6	262042	3	7
9	glucose 6-phosphate isomerase	GI:52782818	<i>C. glabrata</i>	61.3	392050	3	6
36	glucokinase	GI:7434199	<i>A. niger</i>	54.8	293374	3	6
12	dihydrolipoamide DHase (E3)	GI:49648954	<i>Y. lipolytica</i>	53.5	298288	3	3
25	NADP-specific glutamate DHase	GI:118545	<i>A. nidulans</i>	49.8	235090	3	3
10	transketolase	GI:11595598	<i>N. crassa</i>	75.6	712565	3	2
8	myoinositol 1-phosphate synthase	GI:28950116	<i>N. crassa</i>	59.1	390228	2	6
3	phosphoglucomutase	GI:44980859	<i>A. gossypii</i>	63.1	397552	2	5
18	aspartate transaminase	GI:119177955	<i>C. immitis</i>	46.2	333326	2	5
4	H-trans. 2 sector ATP synthase	GI:44985523	<i>A. gossypii</i>	59.0	327050	2	4
28	S-adenosyl methionine synthase	GI:7271000	<i>C. albicans</i>	42.5	210192	2	4
21	isocitrate DHase sub. 2(Leu B)	GI:19111995	<i>S. pombe</i>	41.3	271554	2	3
30	transaldolase	GI:49656881	<i>D. hansenii</i>	36.0	142412	2	3
24	carbonic anhydrase	GI:74662590	<i>N. crassa</i>	32.4	123012	2	2
29	triose phosphate isomerase	GI:9955867	<i>A. oryzae</i>	27.5	152359	2	2
31	proteasome subunit α type	GI:28949952	<i>N. crassa</i>	29.1	213734	2	2
33	3'-phosphoadenosine 5'-phosphatase	GI:28918778	<i>N. crassa</i>	37.2	216202	2	2
35	guanine nucleotide diss. inhibitor	GI:7492946	<i>S. pombe</i>	49.9	261534	2	2
7	catalase-peroxidase	GI:22773423	<i>N. crassa</i>	83.5	416185	1	3
22	CuZn Superoxide dismutase	GI:51702151	<i>A. nidulans</i>	16.0	95198	1	2
14	acetate kinase	GI:74629149	<i>N. crassa</i>	51.6	301214	1	2
8	Threonine synthase THR4	GI:74637378	<i>C. glabrata</i>	57.6	266951	1	1
20	peptidyl-prolyl cis-trans isomerase	GI:74664433	<i>A. niger</i>	23.3	85881	1	1
23	adenylyl sulfate kinase	GI:74614982	<i>N. crassa</i>	23.4	116966	1	1
26	U6 SNRNA-associated sm-like	GI:44890026	<i>A. fumigatus</i>	13.6	43108	1	1
averages					286571	3.359	4.103

^a With gel band indicated in Figure 1 (gel ID no.), protein name, NCBI entry (gi), fungal species originally used for identification, estimated molecular weight by accession sequence, and number of entries in the virtual polymorphism database generated based on the identified protein. The last two columns show the number of peptides that were found in existing database entries (existing) and the number of new peptides matched on the basis of virtual polymorphism (new).

on any resulting matches. In contrast with the MS-Virtmorph method, SNP search programs and PAM refinement modeling allow only single changes per searchable peptide, are implemented within their respective search algorithms, and do not allow database engineering for use with other search programs.

New Peptide Matches Based on Double Substitutions. Among the 120 doubly substituted matches, we were interested in the depth of substitution needed to make the new matches possible. Figure 2b shows the number of matches in which at least one of the two substitutions had a BLOSUM depth of -1 , 0 , 1 , or 2 . A significant number of these new peptide matches relied on at least one substitution at a BLOSUM depth of 0 (41 peptides), but even more (60) required at least one substitution at -1 . Many of the substitutions allowed in the -1 category are also allowed by the SNP table, but all of the doubly substituted peptides are essentially out of reach of current single substitution methods, and the longer peptides are unlikely to have been solved by de novo interpretation.

De Novo: Exact Tag Matching or Homology Searching for Partial Peptide Matches. A set of related search methods that address the problem of matching peptides with sequence varia-

tions, modification, or both exploit de novo interpreted sequences for individual protein or full database searches. Error-tolerant searching as implemented by Mann and Wilm¹³ was the first example in this category and is focused on an individual protein sequence. As mentioned above, discrepancies in peptide mass are reconciled after an exact matching sequence tag is found in the protein sequence, by implying substitutions based on SNP, or modifications. This strategy does not recognize potential matches in which the substitution is contained within the interpreted tag. GutenTag can utilize multiple sequence tags derived from each spectrum, but adds a rigorous quality assessment component with scoring that is based on empirically derived fragmentation modeling.³⁵ The training set for the study included selected high-quality spectra from doubly charged peptides derived from yeast LC tandem-MS data. In the presence of a decoy database, nearly 20% of the peptides matched on the basis of a partial sequence solution were correctly identified by GutenTag. The LC tandem-MS data were from a defined peptide/protein digest mixture, but GutenTag stands out as the only example in the methods mentioned in this

(35) Tabb, D. L.; Saraf, A.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 6415–6421.

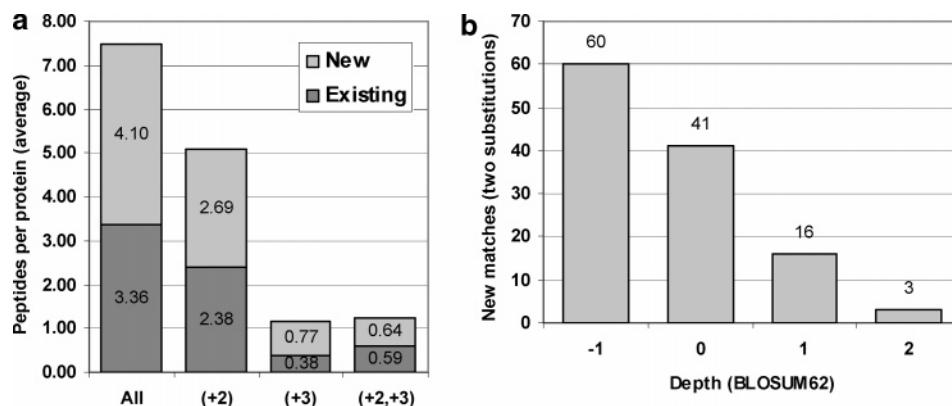


Figure 2. (a) Summary of successful peptide sequence prediction comparing known (existing) peptides found in different species NCBI entries and new peptides (new) based on virtual polymorphism matched in searches. Average number of peptides per protein is shown for each category. All includes peptides isolated in +2, in +3, or in both charge states (+2 and +3), and shown separately for each category as well. Peptide matches in the (+2 and +3) category are based on scans of peptides that were isolated for MS/MS in both charge states independently. (b) New double-substituted peptides, sorted by BLOSUM depth. One of the two substitutions that allowed the match to occur needed to be at the indicated depth (log odds score, rounded to integer²⁰).

article for finding potential substituted/modified sequences that provided performance evaluation in the presence of a decoy database. GutenTag was developed with the goal of full database search capability.

Another distinct approach combines de novo interpretation with homology searching.^{36–38} MS-Blast is one example of such a strategy³⁶ and was tested with complete and partial peptide sequences from spectra interpreted by PredictSequence (Applied Biosystems) from MALDI-Qq-tof MS/MS data. Tags were concatenated with implied gaps (–) to form a query sequence, and BLAST searches were conducted using a modified PAM30 matrix (removing isobaric substitutions, etc). This strategy has the advantage of being applicable as a first pass search and was used to identify a set of proteins from the unsequenced organism *Pichia pastoris*, with resulting improvements based on homology hits. The significance of a potential high scoring pair (HSP; alignment of the query sequence tag and the corresponding database entry segment) was addressed by generating a set of scores from reversed query sequences for comparison. However, the authors suggested that visual inspection of HSPs was often the best way to discriminate between real and false positive alignment HSPs. MS-Blast was proposed as a final resort after peptide mass fingerprinting, and MS/MS searching (Mascot), making this an efficient work flow alternative. A significant amount of advance work in preparing a sequence query may be necessary, and reliance on automated or manual de novo interpretation of spectra for this process introduces an element of risk. Also, most tandem-MS data typically contain at least hundreds of scans from more than one protein, opening the possibility for inadvertently grouping unrelated interpreted sequences from different proteins into a single query. Therefore, like MS-Virtmorph, homology searching of sequence tags by MS-Blast may be limited to samples of low complexity.

Assessing the Potential for Random Matches in MS-Virtmorph Searches.

To provide a probability index of random

events in a database search, the matched peptide can be specifically related to the number of potential matches present in the database. For instance, Mascot⁴ provides a threshold value at a default 95% confidence (user definable, with the default set at a chance of 1 in 20 that the match is a random event) based on the number of peptides in the database that have the same mass (within search tolerance). Sequest provides a similar statistic, P(pep), for assessing individual peptide matches, though it was not available in version 3.1 used for this work. Another useful value to consider is the difference between the best matching peptide and the next best matching peptide in the database (DeltaCn, in Sequest). The MS-Virtmorph databases used in this work, however, have many very similar peptides with similar mass because of substitutions that do not change peptide mass substantially (Glu to Gln, for instance) or as a result of compensatory double substitutions, and the difference between the best match and second best match, DeltaCn, could be very small for related reasons.

Randomized, reversed, or defined decoy databases have been traditionally used to discover actual false positive performance under the experimental conditions used in a study^{35,39} or to reveal performance standards for full database searches so that thresholds can be set to achieve a specific false positive rate.^{40,41} These methods are oriented toward defining false positive rates in fully automated searches where complex samples are often employed and a valid estimate of incorrect assignments is sought. With any error-tolerant search method that allows substitution (including MS-Virtmorph), however, searches must be limited to focusing on individual protein candidates, because of the certainty of obtaining random matches if substitution is implemented on a full database. Moreover, the sample itself cannot be very complex (limits are currently unknown), or one risks the same problem, in reverse. At best, error-tolerant approaches hope to expand the search space (database) with new possibilities in an intuitive manner that predicts the correct peptide sequence along with a minimum number of closely related possibilities. Without excep-

(36) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917–1926.

(37) Mackey, A. J.; Haystead, T. A. J.; Pearson, W. R. *Mol. Cell. Proteomics* **2002**, *1*, 139–147.

(38) Huang, L. H.; Jacob, R. J.; Pegg, S. C.-H.; Baldwin, M. A.; Wang, C. C.; Burlingame, A. L.; Babbitt, P. C. *J. Biol. Chem.* **2001**, *276*, 28327–28339.

(39) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *Omics* **2002**, *6*, 207–212.

(40) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.

(41) Higdson, R.; Hogan, J. M.; Van Belle, G.; Kolker, E. *Omics* **2005**, *9*, 364–378.

Table 2. Forward and Reversed Virtual Polymorphism-Containing Database Search Summary^a

reversed database search results		forward search: highest Xcorr		reverse search: highest Xcorr	
spot	identified protein	(+2)	(+3)	(+2)	(+3)
13	NAD-dependent formate Dase	3.531	7.723	2.634	3.785
9	glucose 6-phosphate isomerase	5.820	6.615	3.298	3.766
36	glucokinase	5.197	5.046	2.771	3.579
22	CuZn superoxide dismutase	4.360	none	2.817	none
28	S-adenosyl methionine synthase	3.614	6.785	2.541	none
8	myoinositol 1-phosphate synthase	5.488	5.078	2.925	3.555
16	sulfate adenylyltransferase	5.404	3.682	2.537	none
3	phosphoglucosmutase	3.621	5.162	3.224	3.732
35	GDP dissociation Inhibitor	3.488	none	2.557	none
15	fumarate lyase	3.538	7.515	2.590	4.143
24	carbonic anhydrase	3.068	3.613	2.450	none
18	aspartate transaminase	3.776	4.663	none	none
25	NADP-specific glutamate Dase	3.844	5.312	2.796	none
32	14-3-3 homologue	3.528	3.994	2.631	none
4	H-trans. 2 sector ATP synthase	4.515	none	2.993	none
7	catalase—peroxidase	5.032	none	3.145	3.558

^a Showing best forward search score (Xcorr) and best reversed search score, by charge state. Searches that failed to find a scan that met the basic scoring criteria are marked “none”.

tion, every method mentioned in this article includes advice to manually validate all matches. This is actually not surprising, since any match that assumes an implied difference from a known reference sequence, for any reason, is a questionable one. Standard approaches for manual interpretation and validation of MS/MS spectra are available.^{24,25}

If correct matches do occur in the MS-Virtmorph database searches, then a search of reversed sequences should result in only the residual level of random matches that should receive distinctly lower scores. This is essentially the same rationale presented for reversing sequence queries for MS-Blast searches and used in the evaluation of HSPs.³⁶ Therefore, we searched a set of databases with all virtual entries reversed to get a sense of the potential for random matches. The heuristic Xcorr score does not vary with the database size, was used for basic threshold criteria in the forward searches, and was therefore utilized for the reversed databases as well. After the search was complete, the basic Xcorr scoring criteria for charge state were applied (see Experimental Section). The reversed database searches were conducted in the absence and then in the presence of the forward MS-Virtmorph database. In each case, all validated and accepted peptides were still identified and presented in the search results when the reversed sequences were present, demonstrating that Sequest could still find and retain them in a substantially more complex database.

Table 2 shows the summary of 16 reversed database searches completed, compared with their corresponding forward searches. The best Xcorr obtained for a peptide match based on the indicated charge state is shown in Table 2. There were four searches that did not receive a single match meeting the basic criteria in the +3 category in the forward searches, whereas there were significantly more in the reverse database searches that had no match meeting the basic criteria; 9 in the +3 category, and 1 in the +2 category. Figure 3 summarizes the average ($\pm$ standard deviation, SD) of the scores in Table 2, as well as the average score of the best matching scan for all accepted peptides found in all forward searches (Table S-1). The best matches in the

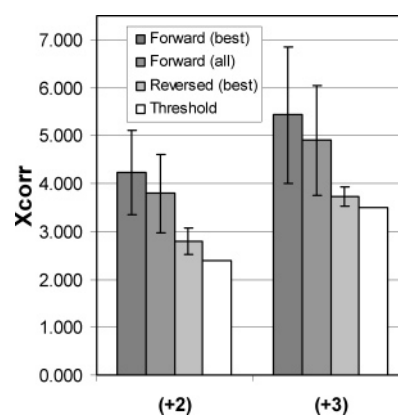


Figure 3. Average of best matches by Sequest Xcorr scores ($\pm$ standard deviation), forward versus reversed database search, by charge state. Categories include the following: (1) best peptide match for each protein found in forward searches in Table 4 (forward, best), (2) best individual score for every accepted peptide match by charge state for all 161 new peptides from Supporting Information Table S-2 (forward, all), (3) highest scoring individual match in the reversed database search in Table 3 (reversed, best), and (4) charge state-based threshold values used for initial filtering (threshold).

reversed database searches (reversed best) were, on average, just above the threshold criteria, whereas the best Xcorr values in the forward searches (forward best) were higher. The average value (best matching scan) for all 160 accepted peptide matches in the +2 and +3 categories (forward all) were also higher on average, though overlapped (SD) with the best reversed matches. On the basis of this analysis, it might seem useful to increase the basic threshold criteria, but many legitimate peptide matches can score just above these threshold Xcorr values. This analysis provides good evidence that automated searching of these databases resulted in the discovery of new matches not attributable to random events.

It is conceivable that the sequence of a semitryptic, nontryptic, or modified peptide represented in an MS/MS spectrum may align correctly with only the C-terminal or N-terminal sequence of a virtual entry that coincidentally has the same mass within search

Table 3. New Peptides Found for Proteins Identified on the Basis of One or Two Peptide Matches from Existing NCBI Entries.

Proteins with One Existing Peptide Match	
average protein size, 38.43 (kDa)	<i>new</i>
average number of new peptides/protein,	1.57
adenylate kinase (23.4)	1
THR4 threonine synthase (57.6)	1
CuZn superoxide dismutase (16.0)	2
peptidyl-prolylCT isomerase (23.3)	1
U6 SNRNA-associated sm-like protein (13.6)	1
catalase-peroxidase (83.5)	3
acetate kinase (51.6)	2
Proteins with Two Existing Peptide Matches	
average protein size: 43.63 (kDa)	<i>new</i>
average number of new peptides/protein:	3.33
triose phosphate isomerase (27.5)	2
transaldolase (36.0)	3
myoinositol 1-phosphate synthase (59.1)	6
GDP diss. inhibitor (49.9)	2
H-trans. 2 sector ATP synthase (59.0)	4
S-adenosyl methionine synthase (42.5)	4
3'-phosphoadenosine 5'-phosphatase (37.2)	2
phosphoglucosyltransferase (63.1)	5
carbonic anhydrase (32.4)	2
aspartate transaminase (46.2)	5
isocitrate dehydrogenase subunit 2 (Leu B) (41.3)	3
proteasome subunit α type (29.4)	2

tolerance. Any match of this nature would not be entirely random, may receive a good Xcorr score, may withstand validation, and therefore may inadvertently enhance the quality of a protein identification. This type of false positive might occur in any error-tolerant search approach, but it is not possible to predict or easily detect such events, until the corresponding gene has been sequenced.

Improving Marginal Matches. The most useful application of this method may be realized in the improvement of marginally significant protein identifications. Investigators make difficult decisions on whether to invest time and resources on identifications in this category, and any additional evidence to support the identification can be critical. Table 3 shows new matches gained for proteins initially identified on the basis of one or two peptides. In this limited data set, seven proteins with one existing peptide match obtained 1.57 new matches on average, whereas proteins with two existing matches gained an average of 3.33 new matches. With a few notable exceptions, there is some correlation of the average size of these proteins (38.3 versus 43.6 kDa) with the number of existing matches, and with the number of new matches, which would not be unexpected. In this more difficult category of marginal matches, one might also compare the sequence of the matched virtual peptide to the corresponding region in a multiple sequence alignment of the protein in a divergent set of species to evaluate its potential for biological compatibility.

CONCLUSIONS AND PERSPECTIVE

Database matching search programs effectively identify proteins by comparing known, existing database entry-derived virtual data to experimental MS/MS spectral data. When difficulties are encountered in identifying proteins from the more divergent species that have not been sequenced, de novo interpretation and error-tolerant strategies provide useful second pass methods for improving identifications by presenting additional potential peptide

matches. Methods that combine the two strategies are more powerful, but still cannot routinely identify peptides from more divergent proteins, due to the common occurrence of mass error and absent or extraneous peaks. Database searching with sequence tags, and hybrid methods that make use of homology searching, can also effectively bypass the need for exact matches but depend on de novo interpretation. In contrast, MS-Virtmorph databases exploit the strength of database searching programs for their ability to find matches without the need for comprehensive fragment coverage or accurate mass. The alternative methods addressed in this article are not comprehensive by any means, but we found no other example that focuses on improving the prediction of polymorphism and substitution, with broad compatibility across platforms. This database engineering strategy allows flexibility in engineering sequences to be considered and can predict both minor and more complex divergence. As with other error-tolerant approaches, it is limited as a second pass strategy based on a preliminary protein match (not scalable for shotgun protein identification), and it is not recommended for samples of high protein complexity. This strategy can be immediately implemented by those facing the identification of proteins in unusual species, offering a relatively simple way to use their existing database search programs to find potential new peptide matches to improve protein identifications. It can also be applied retroactively to archives of data that were set aside after exhausting methods available at the time.

We chose BLOSUM62 for the current work because it offers the potential for spanning species divergence for the problem of protein identification in *A. Alternata* and because assembling an amino acid substitution program based on a generic matrix was easily accomplished. A more restrictive generic matrix such as BLOSUM90 (from sequence blocks that are no less than 90% identical), for instance, may be useful in predicting polymorphism within species more effectively than by SNP prediction. Alternatively, one could choose different matrices at different substitution levels to restrict the size of a database while including moderate divergence. However, some amino acid positions in a given protein have strict conservation for biological reasons that cannot be predicted by any generic amino acid substitution matrix. Moreover, overall sequence conservation can be very high, or very low (histones versus albumin, for example), requiring very different substitution prediction. Ideally, the placement of virtual polymorphism should be based on the characteristics of individual proteins or protein families. A position-specific substitution matrix like those generated and used for finding related proteins in a database (PSI-BLAST⁴²) could provide a more restrictive, highly relevant substitution matrix for an individual protein. Such a matrix would naturally provide automated position-specific exclusion of invariant amino acid residues and would limit substitutions in an uncompromising, biologically relevant manner. This would substantially reduce the number of potential sequences engineered, restrict database size, and reduce search times. Moreover, since all peptides generated would be based on substitution observed in at least one aligned protein member, the probability of considering a biologically irrelevant sequence would be minimized. Such an algorithm will be considered in future work.

(42) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

ACKNOWLEDGMENT

The authors thank Dr. Micheal Kinter, Ph.D. (Cleveland Clinic Foundation, Lerner Research Institute, Cleveland, OH) for review and advisement regarding this paper, and Ms. Ada Solidar (Wyckoff Lab) for help and advice in the implementation of this method.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review February 19, 2007. Accepted April 18, 2007.

AC0703496