Anal. Chem. **2001,** *73,* 1917−1926

Accelerated Articles

Charting the Proteomes of Organisms with Unsequenced Genomes by MALDI-Quadrupole Time-of-Flight Mass Spectrometry and BLAST Homology Searching

Andrej Shevchenko,*[,†,‡] Shamil Sunyaev,[§] Alexander Loboda,[||,⊥] Anna Shevchenko,[†,‡] Peer Bork,[§] Werner Ens,[⊥] and Kenneth G. Standing[⊥]

Peptide and Protein Group, European Molecular Biology Laboratory (EMBL), 69012 Heidelberg, Germany, Structural and Computational Biology Program, European Molecular Biology Laboratory (EMBL), 69012 Heidelberg, Germany, MDS Sciex, 71 Four Valley Drive, Concord ON L4K 4V8, Canada, and Department of Physics and Astronomy, University of Manitoba, Winnipeg MB R3T 2N2, Canada

MALDI-quadrupole time-of-flight mass spectrometry was applied to identify proteins from organisms whose genomes are still unknown. The identification was carried out by successively searching a sequence database—first with a peptide mass fingerprint, then with a packet of noninterpreted MS/MS spectra, and finally with peptide sequences obtained by automated interpretation of the MS/MS spectra. A "MS BLAST" homology searching protocol was developed to overcome specific limitations imposed by mass spectrometric data, such as the limited accuracy of de novo sequence predictions. This approach was tested in a small-scale proteomic project involving the identification of 15 bands of gel-separated proteins from the methylotrophic yeast *Pichia pastoris*, whose genome has not yet been sequenced and which is only distantly related to other fungi.

Mass spectrometry has been widely recognized as a cornerstone of proteomic research because of its high sensitivity and throughput (reviewed in refs 1−3). Proteins separated by one-dimensional or two-dimensional gel electrophoresis can be digested in-gel and rapidly identified at the femtomole level by MALDI peptide mapping, by tandem mass spectrometry, or by a combination of those techniques (reviewed in refs 2 and 4−6). Alternatively, unfractionated mixtures of proteins isolated in biochemical experiments can be enzymatically digested in-solution, followed by peptide sequencing by LC MS/MS[7,8] or CIE MS/MS.[9] Although protein identification may be achieved by a large variety of mass spectrometric techniques, it ultimately requires that the acquired mass spectra be accurately matched to protein sequences from the corresponding database entries. Therefore, the availability of a complete genome or at least a substantial part of the cDNA sequences is of paramount importance.[2,10,11]

What if the protein of interest is not present in a database? The simplest case is when the sequence of a highly homologous protein from another species is available. Enzymatic digestion of the protein of interest would then be expected to yield some

* Corresponding author: (fax) +49 6221 387 306; (e-mail) shevchenko@EMBL-Heidelberg.de.

† Peptide and Protein Group, EMBL.

‡ Present address: MPI of Molecular Cell Biology and Genetics, 01307 Dresden, Germany.

§ Structural and Computational Biology Program, EMBL.

|| MDS Sciex.

⊥ University of Manitoba.

(1) Blackstock, W. P.; Weir, M. P. *Trends Biotechnol.* **1999**, *17*, 121−127.
(2) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837−846.
(3) Neubauer, G.; Wilm, M. *Curr. Opin. Mol. Ther.* **1999**, *1*, 695−701.
(4) Yates, J. R. *J. Mass Spectrom.* **1998**, *33*, 1−19.
(5) Chalmers, M. J.; Gaskell, S. J. *Curr. Opin. Biotechnol.* **2000**, *11*, 384−390.
(6) Lahm, H. W.; Langen, H. *Electrophoresis* **2000**, *21*, 2105−2114.
(7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd *Nat. Biotechnol.* **1999**, *17*, 676−682.
(8) Martin, S. E.; Shabanowitz, J.; Hunt, D. F.; Marto, J. A. *Anal. Chem.* **2000**, *72*, 4266−4274.
(9) Jensen, P. K.; Paša-Tolić, L.; Anderson, G. A.; Horner, J. A.; Lipton, M. S.; Bruce, J. E.; Smith, R. D. *Anal. Chem.* **1999**, *71*, 2076−2084.
(10) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440−14445.
(11) Pandey, A.; Lewitter, F. *Trends Biochem. Sci.* **1999**, *24*, 276−280.

10.1021/ac0013709 CCC: $20.00
Published on Web 03/23/2001

peptides identical to the ones present in the known protein homologue. If the number of identical peptide masses is sufficient to produce a statistically reliable hit upon database searching, such proteins can often be identified by MALDI peptide mapping.[12,13] However if the similarity between the proteins is so low that only a few identical peptide sequences are shared, the next step is sequencing by tandem mass spectrometry. In principle, tandem mass spectrometry can identify the protein by a single matching peptide[14,15] or by error-tolerant database searching.[13,16−17] Although a number of proteins have been identified in this way, the approach typically requires manual interpretation of MS/MS data and careful inspection of the match.[18] Moreover, error-tolerant searching using peptide sequence tags is unlikely to hit peptides that differ by multiple amino acid substitutions from relevant sequences in a database.

Sequence similarity searches have been also employed to identify proteins via their known homologues in other species. The CIDentify program developed by Taylor and Johnson[19] uses a modified FASTA sequence comparison algorithm to screen the sequences produced by automated interpretation of low-energy CID spectra. However, because of rapid growth of sequence databases, the throughput of the approach is limited by the relatively long running time required by the FASTA algorithm. A much faster alternative is offered by the BLAST searching tool.[20] BLAST effectively identifies alignment "seeds" and then extends alignments around the seed. Therefore, an overwhelming majority of database sequences is discarded without aligning with the queried sequence thus dramatically decreasing the algorithm's running time. Advanced BLAST programs are also operated at servers of very high computational capacity that are accessible over the Web.[21] However, conventional BLAST searching at most of those servers is optimized to identify similarities between fairly long protein sequences and therefore has limited value for screening peptide sequences produced by mass spectrometry.

If sequence similarity is still too low, or if the gene is new, tryptic peptides must be sequenced de novo.[22,23] Peptide sequences are used for designing oligonucleotide probes, and the full length sequence of the corresponding protein is subsequently determined via cloning of the cognate gene by a PCR-based approach.[24] Although the combination of a hybrid quadrupole time-of-flight instrument and C-terminal isotopic labeling of peptides[25] together with "differential scanning"[26] has facilitated de novo sequencing, it still remains laborious and time-consuming. Moreover, subsequent cloning presents even more technical challenges, so such an approach has so far not been applied in proteomic projects.

On the other hand, genomic sequencing has made truly spectacular progress in recent years. More than 30 prokaryotic genomes are publicly available (see, e.g., http://www.tigr.org/tdb/mdb/mdbcomplete.html) as well as genomes of eukaryotic organisms such as *Saccharomyces cerevisiae*,[27] *Caenorhabditis elegans*,[28] and *Drosophila melanogaster*.[29] The human genome is scheduled for completion by the year 2003[30,31] with a draft (∼90% of a consecutive sequence) already available.[32,33] Thus, given the size and completeness of the sequence databases, it is conceivable that many proteins from an organism with an unknown genome are likely to have homologues already present in a database. If so, mass spectrometry might provide sufficient information to identify such a homologue, and accurate de novo sequencing followed by PCR-based cloning would be required only in exceptional cases.

It would therefore be a significant advantage if a single mass spectrometric experiment could produce data for identifying proteins by peptide mass mapping, tandem mass spectrometric sequencing, and homology searching. Various database mining strategies could then be applied consecutively, starting from peptide mass mapping for the most straightforward cross-species identifications and ending up with automated interpretation of MS/MS spectra and homology searching.

We therefore set out to investigate whether the recently introduced technique of MALDI-quadrupole TOF mass spectrometry (MALDI-QqTOF MS)[34−36] could meet those requirements. Indeed, MALDI-QqTOF mass spectrometers enable a peptide mass map and a number of tandem mass spectra from selected peptide precursors to be acquired in a single experiment. More than 10 000 resolution (fwhm) and better than 20 ppm mass accuracy in both MS and MS/MS modes allow very specific database searching. Although MS/MS spectra acquired on a MALDI-QqTOF mass spectrometer do not usually contain continuous series of y- or b-ions, the peptide sequences can be predicted by available software with reasonable accuracy.[35]

(12) Podtelejnikov, A. V.; Bachi, A.; Mann, M. *Proc. 46th ASMS Conf. Mass Spectrom. Allied Top*., Orlando, FL, 1998; p 212.
(13) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871−2882.
(14) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466−469.
(15) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.
(16) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(17) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290−299.
(18) Shevchenko, A.; Keller, P.; Scheiffele, P.; Mann, M.; Simons, K. *Electrophoresis* **1997**, *18*, 2591−2600.
(19) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(20) Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(21) Gaeta, B. A. *Biotechniques* **2000**, *28*, 436−440.
(22) Shevchenko, A.; Wilm, M.; Mann, M. *J. Protein Chem.* **1997**, *16*, 481−490.
(23) Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. In *Protein in Peptide Analysis*; Chapman, J. R., Ed.; Humana Press: Totowa, NJ, 2000; Vol. 146, pp 1−16.
(24) Lingner, J.; Hughes, T. R.; Shevchenko, A.; Mann, M.; Lundblad, V.; Cech, T. R. *Science* **1997**, *276*, 561−567.
(25) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1015−1024.
(26) Wilm, M.; Neubauer, G.; L., T.; Shevchenko, A.; Bachi, A. In *Proteome and protein analysis*; Kamp, R. M., Kyriakidis, D., Choli-Papadopoulou, T., Eds.; Springer: Berlin, Heidelberg, New York, 1999; pp 65−79.
(27) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. B. *Science* **1996**, *274*, 546−567.
(28) The *C. elegans* Sequencing Consortium. *Science* **1998**, *282*, 2012−2018.
(29) Adams, M. D.; et al. *Science* **2000**, *287*, 2185−2195.
(30) Waterston, R.; Sulston, J. E. *Science* **1998**, *282*, 53−54.
(31) Collins, F. S.; Patrinos, A.; Jordan, E.; Chakravarti, A.; Gesteland, R.; Walters, L. *Science* **1998**, *282*, 682−689.
(32) International Human Genome Sequencing Consortium. *Nature* **2001**, *409*, 860−921.
(33) Venter, J. C.; et al. *Science* **2001**, *291*, 1304−1351.
(34) Shevchenko, A.; Loboda, A.; Shevchenko, A.; Ens, W.; Standing, K. G. *Anal. Chem.* **2000**, *72*, 2132−2141.
(35) Loboda, A. V.; Krutchinsky, A. N.; Bromirski, M.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1047−1057.
(36) Krutchinsky, A. N.; Zhang, W.; Chait, B. T. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 493−504.

We here present a strategy for identifying homologous proteins by a combination of MALDI-QqTOF mass spectrometry and a modified BLAST searching protocol that we call "MS BLAST" (mass spectrometry-driven BLAST searching). We tested the approach by identification of 15 gel-separated proteins from the yeast *Pichia pastoris* whose full sequences have not yet been determined.

EXPERIMENTAL SECTION

Materials and Reagents. Unless otherwise noted, all chemicals were purchased from Sigma (Sigma Chemicals, St. Louis, MO) and were of analytical grade. For mass spectrometric analysis and preparation of digests, HPLC grade water, methanol, and acetonitrile (LabScan, Dublin, Ireland) were used.

A preparation of membrane proteins from the MDCK (Madin Darby Canine Kidney) cell line was obtained from Dr. K. Simons's laboratory (Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany). Proteins were purified by flotation and one-dimensional gel electrophoresis and were visualized by Coomassie staining.

Proteins from *P. pastoris* were isolated in Dr. A. Hyman's laboratory (Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany). The cell extract was first fractionated on a MonoQ ion exchange column (Pharmacia Amersham). Fractions were further analyzed by one-dimensional gel electrophoresis, and proteins were visualized by Coomassie staining.

In-Gel Digestion and Preparation of Sample Probes. Proteins were in-gel digested with trypsin (unmodified, sequencing grade, Roche Diagnostics GmbH, Mannheim, Germany) as described previously.[15,23]

Tryptic peptides were extracted from a gel matrix with 5% formic acid and acetonitrile; the extracts were pooled and dried in a vacuum centrifuge. The digests were then redissolved in 5 μL of 5% formic acid. Typically a 0.5-μL aliquot was deposited on top of a spot of 2,5-dihydroxybenzoic acid (DHB) (Sigma Chemicals) matrix prepared as described.[34]

Acquisition of MS and MS/MS Spectra. All experiments were performed on a prototype MALDI-QqTOF mass spectrometer built at the University of Manitoba in collaboration with MDS Sciex (Concord, ON, Canada).[35] The collision energy was set by applying an initial accelerating voltage at the entrance of the collision cell. The energy was chosen according to the rule 0.05 V/Da with further tuning if necessary. The instrument was calibrated externally, and no postacquisition recalibration of MS and MS/MS spectra was performed.

Data Interpretation and Database Searching. Searching with all types of data was performed against a comprehensive nonredundant protein sequence database. No limitations on protein molecular weights, p*I*, or species of origin were imposed.

Database searching using a peptide mass map was performed using the PeptideSearch v 3.0 program.[37]

Database searching using tandem mass spectra was performed at the Matrix Science Ltd. server (http://www.matrixscience.com/) using MASCOT software.[38] Noise in the spectra was filtered out using Microsoft Excel (Microsoft Corp.), and the spectra were submitted for database searching in a generic MASCOT format. Mass tolerance was set at 50 ppm for the masses of peptide precursors and at 0.05 Da for the masses of fragment ions.

The PredictSequence routine (a part of the BioMultiview 1.4 software, MDS Sciex) was used for de novo sequence interpretation of tandem mass spectra. Where specified, the BioTools (a part of the Analyst 1.0 alpha 4 software, MDS Sciex) was applied.

MS BLAST Homology Searching Protocol. All complete and partial peptide sequences obtained by PredictSequence were used for MS BLAST searching. Sequences were edited according to the following rules:

(a) L stands for both leucine and isoleucine residues. Z stands for glutamine and lysine residues, which occur elsewhere in the peptide sequence, and is used if reliable differential assignment of these amino acid residues in the peptide sequences is not possible. K stands for a C-terminal lysine residue.

(b) If the peptide sequence is complete (i.e., the calculated mass fits the mass of the precursor ion), a symbol of the trypsin cleavage site B is inserted prior to the sequence.

(c) The symbol X stands for an undefined amino acid residue.

(d) All sequence proposals obtained for sequenced peptides are spaced with the minus symbol (−) and are merged into a single string. The query may contain space symbols, hard returns, numbers, etc., all of which will be ignored by the server. We note here that the scoring values for B and X symbols in the conventional PAM30 matrix[20] are substituted by the newly defined scores as explained below.

Searching was performed by WU-BLAST2 program (Gish, W. (1996−1999) http://blast.wustl.edu) provided at the EMBL server: http://dove.embl-heidelberg.de/Blast2/. The following settings were applied: "Program", blast2p; "Database", nrdb95; "Matrix", PAM30MS; "Expect", 100; "Other advanced options", -nogap -hspmax 100 -sort_by_totalscore -span1.

The algorithms of BLAST homology searching, statistical analysis, and organization of the search engine are described in refs 20 and 39. Practical guidance on BLAST searching and explanation of settings and available options are provided at the Web pages of most of BLAST servers.[21]

RESULTS AND DISCUSSION

A protein identification strategy for analysis of homologous proteins is outlined in Figure 1. First, a peptide mass map is acquired and a database is searched with a set of accurately determined peptide masses. Mass fingerprinting applied as a first "screening" step enables rapid identification of known proteins and proteins highly homologous to them. If a plausible protein candidate has been hit, the match can be verified further by tandem mass spectrometric investigation of the selected *matched* peaks.[34]

If no protein candidate has been hit or if a number of intense peaks that do not match the already identified protein have been detected, tandem mass spectra are subsequently acquired from *as many unknown peptides as possible*. Noninterpreted spectra are submitted in a single packet for searching by MASCOT. Within minutes, the program identifies the peptides identical to the ones

(37) Mann, M. In *Microcharacterization of Proteins*; Kellner, R., Lottspeich, F., Meyer, H. E., Eds.; VCH: Weinheim, 1994; pp 223−245.

(38) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

(39) Altschul, S. F.; Boguski, M. S.; Gish, W.; Wootton, J. C. *Nat. Genet.* **1994**, *6*, 119−129.

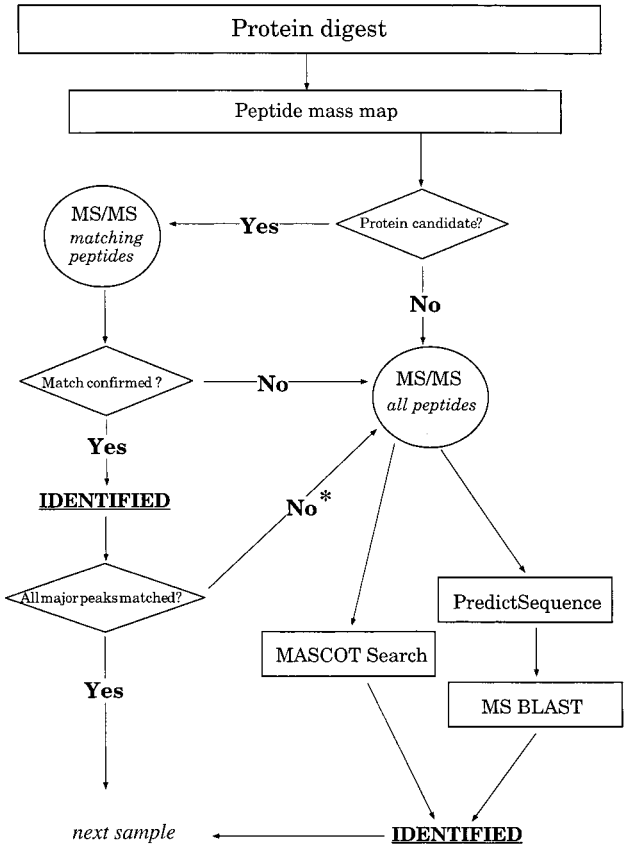

Figure 1. A strategy for MALDI-QqTOF identification of proteins isolated from organisms whose genome has not been sequenced. (*) refers only to those peaks that do not match the already identified protein(s).

in a database. One or two peptides matched with statistically significant scores produce a strong hit, leading to unambiguous identification of a homologous protein.

If, however, the MASCOT score is statistically unreliable, all acquired MS/MS spectra are further interpreted de novo by the PredictSequence software; then the resulting sequence proposals are merged and submitted for MS BLAST searching. As demonstrated below, MS BLAST can match similar peptide sequences and either identify the protein or confirm a vague hit produced by MASCOT.

We note here that the approach utilizes only automated processing of data. No manual interpretation of tandem spectra or error-tolerant database searching is involved, and consequently, the whole strategy could be completely automated. It thus has the potential to become a high-throughput tool for protein identification.

Identifying Proteins by the MS BLAST Protocol. We developed a BLAST searching protocol (MS BLAST) for identification of proteins by homology searching using peptide sequences produced by mass spectrometry. The concept of the BLAST algorithm[40] ideally suits this task. Ungapped BLAST search identifies all high-scoring pairs (HSPs)—regions of high local sequence similarity—between individual peptides in the query and a protein sequence from the database entry. Importantly, the sequential order of the matched segments does not affect the total score, which is calculated for each protein entry by adding up the scores of individual HSPs that are higher than the specified threshold. Therefore, in the MS BLAST protocol, all peptide sequence proposals obtained by the interpretation of all MS/MS spectra are merged in an arbitrary order into a "chimeric sequence" and a single database search is performed. In the query, peptide sequences are spaced with the gap symbol (−) to which a high negative score has been assigned. This prevents the algorithm from reporting false similarities with subsequences involving parts of peptide sequences adjacent in a query string.

MS BLAST is targeted at matching of closely related relatively short peptide sequences. Therefore, gaps are not permitted and a substitution matrix that is used for calculating the scores of individual HSPs is adjusted to reporting of aligned sequences of high similarity. At the same time, a diagonal scoring matrix that strongly favors the identity of matched sequences cannot be used because of possible amino acid substitutions and/or numerous errors in peptide sequence proposals. Therefore, we introduced several modifications into the PAM30 scoring matrix, which accommodate specific requirements imposed by MS/MS sequencing:

(i) Scores for the isobaric amino acids leucine/isoleucine and glutamine/lysine were substituted for their average values.

(ii) The specificity of trypsin was considered by reserving the K symbol for a C-terminal lysine and by introducing the symbol B with a value averaged between arginine and lysine to represent a cleavage site preceding the peptide sequence.

(iii) The symbol X, standing for an undefined amino acid residue, was introduced with zero scoring value for any substitution. Since BLAST penalizes insertions/deletions in aligned sequences, the use of X and B symbols increases the score if a queried peptide sequence is spaced from the cleavage site by a specified number of amino acid residues.

Several important options of MS BLAST are specified via the command line. MS BLAST sorts the retrieved hits by the total score of HSPs. This brings to the top of the list those proteins that have been matched to multiple peptide sequences, even though the score for any individual HSP might be rather low. Automated interpretation of MS/MS spectra produces many similar, although distinct versions of the peptide sequence for the same spectra. Those similar sequences may match the same region of the protein sequence in a database, thus resulting in an erroneously high total score. Command line option "span1" helps to overcome this complication by skipping HSPs spanned on query, subject, or both by a better HSP. Therefore, by selecting the best matching HSP from a number of redundant sequences, and by sorting the hits according to their total scores, MS BLAST effectively fulfills the task of a result compiler in the course of database searching.

In the case study presented below, we applied a combination of MALDI-QqTOF and MS BLAST to identify a 50 kDa protein isolated from canine cells. A peptide mass map acquired from an in-gel digest of the band is presented in Figure 2A. Although the masses of 35 peptide ions were used for database searching with 50 ppm mass tolerance, no hit was produced. Subsequently, tandem mass spectra were acquired from 15 of the most intense ions and uploaded into MASCOT. The search yielded with

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

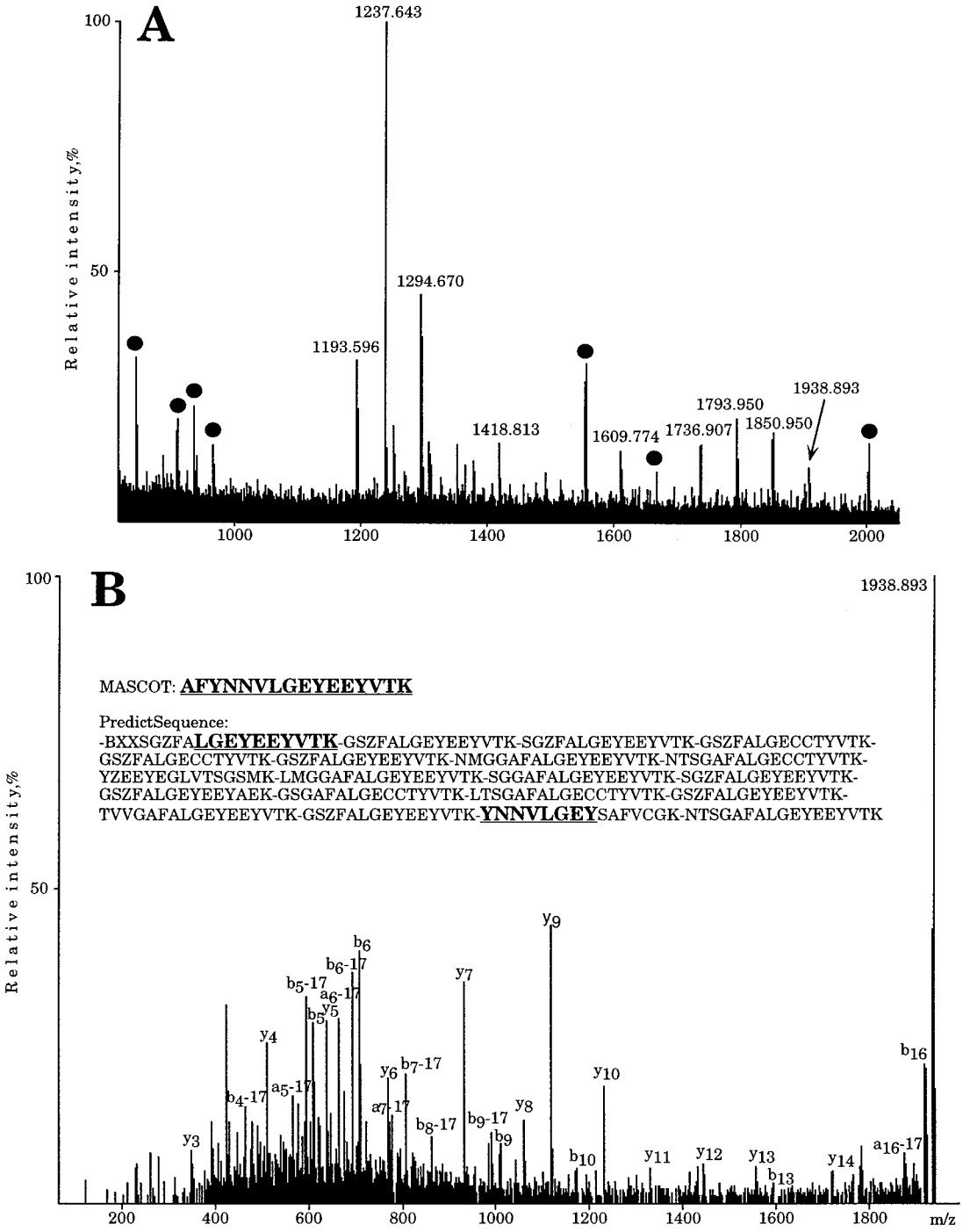

Figure 2. (A) Peptide mass map acquired from the in-gel digest of a 50 kDa canine protein. Peptide precursors from which MS/MS spectra were acquired are designated with filled circles or *m*/*z*. The latter are specified only for the peptide peaks that were matched to OATs by MASCOT or by MS BLAST (listed in Table 1). (B) Tandem mass spectrum acquired from the precursor ion *m*/*z* 1938.893. The sequence of the MASCOT hit and sequences determined by PredictSequence are presented above the spectrum. Sequence stretches matched by MS BLAST are highlighted. Major ion series in the MS/MS spectrum are labeled accordingly to Biemann's nomenclature.

statistically significant scores two peptides from human ornitine aminotransferase (OAT) and one peptide from its *Caenorhabditis elegans* homologue (Table 1). The molecular weight of both OATs (~49 kDa) was close to the apparent molecular weight of the band, strongly suggesting that the band is yet unknown canine OAT. However, only 3 out of 15 MS/MS spectra were assigned. To establish the identity of the other peptide peaks, PredictSequence was run on all 15 MS/MS spectra. A query string comprising 324 peptide sequences in total was assembled (Figure 3), and the MS BLAST search was completed in less than 2 min. OATs from various species occupied the top nine positions in the list of MS BLAST hits. In total, five peptide sequences matched the mouse OAT, and another two short sequences were matched to homologous proteins from *Drosophila* species (Table 1). The precursor ion with *m*/*z* 1294.67 (Figure 3) matched the same peptide from the OAT sequence as *m*/*z* 1237.65 and was not included in Table 1. Altogether, nine peptide ions were matched to various OATs, providing a much more confident identification

Table 1. Identification of the Canine Homologue of Ornitine Aminotransferase (OAT)

m/z	Identified by MASCOT		Identified by MS BLAST		
	Sequence	Org.[a]	High Scoring Pairs	Score	Org.[a]
1237.65	IVFAAGNFWGR	*HS*	Query:1196 FAAGNFWGR 1204 Sbjct: 172 FADGNFWGR 180	62	*MM*
1736.91	YGAHNYHPLPVALER	*HS*	n.i.[b,c]		
1938.89	AFYNNVLGEYEEYVTK	*CE*	Query:4504 YNNVLGEY 4511 4527 LGEYEEYVTK 4536 Sbjct: 116 YNNVLGEY 123 120 LGEYEEYITK 129	62; 71	*MM*
1193.60	n.i.		Query:846 BXXYVWDVEGR 856 Sbjct: 66 KGIYMWDVEGR 76	64	*MM*
1609.77	n.i.		Query:2165 EEGXXSSDYLFER 2177 Sbjct: 34 EQGPPSSEYIFER 46	69	*MM*
1793.95	n.i.		Query:3310 NYHPLEPXXER 3320 Sbjct: 54 NYHPLPVALER 64	45	*MM*
1418.81	n.i.		Query:1622 PZLVAEL 1628 Sbjct: 89 PKIVAAL 95	33	*DM*[d]
1850.95	n.i.		Query:3485 ETGAPA 3490 Sbjct: 25 ETAAPA 30	32	*DA*[d]

[a] Org, organism. Names and abrreivations of organisms are listed at the end of the text. [b] n.i., not identified. [c] PredictSequence did not suggest any sequences similar to the one that was hit by MASCOT. [d] Only the peptides that were not matched also to the top-scoring *MM* protein are presented.

of the canine protein.

The data from Table 1 highlight several important features of MS BLAST. First, MS BLAST can match "islets" of correct sequences rather than sequences of complete peptides only. Comparison with the three peptides matched by MASCOT suggests that none of the corresponding sequences matched by MS BLAST was complete and correct. On one occasion (*m*/*z* 1938.89; Figure 2B), the correct N-terminal stretch was linked to the wrong C-terminal sequence and vice versa. Because of the limited accuracy of automated interpretation of MS/MS spectra, as many sequence proposals as possible need to be included in a query, so 324 candidates were uploaded, and the specificity of MS BLAST was sufficient to fish out 9 partially matching peptides (Figure 3 and Table 1). The candidate sequences produced by the PredictSequence program for each MS/MS spectrum are listed in Figure 3 in order of their relative scores. Note that only three out of the nine matched sequences occupied top positions in the respective lists. Thus, MS BLAST readily overcomes the deficiencies of the sequence prediction software.

We note here that MASCOT and MS BLAST may be pointing to different protein homologues. MASCOT identifies and lists proteins in which peptides exactly matched the uploaded MS/MS spectra. In contrast, MS BLAST sorts hits by their total scores, and therefore, it is sensitive to the total number of HSPs. Consequently, an HSP with a lower score might be included if for the particular protein the score calculated for all HSPs is higher. To identify the best matching HSP for a particular peptide, the search can be repeated with the results sorted by the highest score rather than by the total score.

Thus, we have demonstrated that MS BLAST efficiently complements MASCOT searching by finding additional matching peptides. Indeed, MS BLAST can score a hit even if no identical peptides have been detected at all and the larger number of peptide sequences included in the query compensates the limited accuracy of sequence predictions. At the same time, data interpretation throughput is not degraded, since the search is performed on a server with high computational capacity.

Statistical Evaluation of MS BLAST Hits. Visual inspection of the alignments is often the most reliable way to discriminate between true and false positives. For the MS BLAST-based protein identification, the evaluation of statistical significance of hits is very important because sequences of tryptic peptides are short and because the automated interpretation of MS/MS spectra is extremely prone to errors. However, the standard statistical approach used for ungapped BLAST searches[41] does not adequately meet the above-mentioned specifics of typical MS BLAST queries. Therefore, we suggest setting the significance thresholds conditionally on a number of HSPs reported by MS BLAST for the protein hit and determining those threshold values via computer simulation experiments.

To preserve local amino acid composition of peptides in computer simulations, we obtained random peptide sequences by reversing the sequences of real peptides (basically, by reading them from the C-terminus to the N-terminus). Since MS BLAST engages the "span1" option, the redundant sequences are usually filtered out and only the best matching sequence is reported (see

(41) Karlin, S.; Altschul, S. F. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5873–5877.

908.446 -GVPLFRR-GVPLGMDR-GVMSTPPR-GVPSTMPR-GVLPGMDR-RPLGMDR-GVNANVDR-GVMXXLDR-GVLGPMDR-GVPSTLDR-RPSTMPR-RLPGMDR- GVPSDVDR-
GVMXXMPR-RNANVDR-GDPPGMDR-RLGPMDR-RPSTLDR
935.478 -FGTTDLR-GFTTDLR –TLTDTRR-SGYLGCLR-GSYLGCLR-AAYLTNDR-NLTDTRR-TRTTDLR-SGYLTNDR- SGYLTNDR-TLDTTRR-AAZTTDLR-EDADTRR-
SGYNTDLR-NRTTDLR-AAYLGCLR-EDWTRR-EZTTDLR-ZZTTDLR-NLDTTRR-LATTDLR-ALTTDLR-TLAFVRR-AAYNTDLR
968.614 -HAVALDPR-HAVALPDR-HAVALVLR-HAVALLVR-TAVALXXR -PGAVALXXR-HASNAPPGK-GDAALANHK-PGASNAPPGK-HAVLAXXR-TASNAPPGK-SAALANHK-
APGSNANHK-VGNLAVLR-GPGNLAVLR-HGNLAVLR-TAVLAXXR-TGNLAXXR-PGAVLAXXR-VGNLAXXR -GPGNLAXXR-APGVLAVLR-AVGPLAVLR
1237.643 -BDPFAAGNFWGR-BDPFAAGNFVSGR-BDMLAAGNFWGR-BXXEAAGNFWGR-BDPFAAGNFZDR-BXXGTAAGNFZACP-BXXGTAAGNFVSGR-BDMLAAGNFVSGR-
BXXGZAAGNFWGR-BETEAAGNFVSGR-BEETAAGNFWGR-BXXMAAGNFWGR-BDDEAAGNFVSGR-BXXXXAAGNFZDR-BZCZAAGNFWGR-BDPFAZNFVSGR-
BDMLAAGNFZDR-BXXGZAAGNFVSGR-BDPFAAGNFXXK-BDPFAZNFWGR
1193.596 -BXXYVWDVEGR-BXSYVWDVEGR-SYVDVSVEGR-SYVSVDNGGGR-SYVWDGNGGR-BXXDFWDVEGR-BXXMMWDVEGR-BXGDVVWDVEGR-BXSDFWDVEGR-
VVWDVEGR-BGLTVWDVEGR-AVDVSVEGR-BSYVSVDXXGR-PAYVWDVWR-APYVSVDVWR-TVWDVEGR-YFWDVEGR-SDFSVDXXGGR-YVWDVZLT-
BXXGVDVSVEGR
1294.670 -PAGFAAGNFVSGR-APGFAAGNFVSGR-GGFAAGNFVSGR-PGLFAAGNFVSGR- GPLFAAGNFVSGR-LPGFAAGNFVSGR-PLGFAAGNFVSGR-VPGFAAGNFVSGR-
PVGFAAGNFVSGR-GGFAAGNFWGR-NFAAGNFVSGR-MFAAGNFVSGR-RLFAAGNFVSGR-GGFAAZFAASGR-GFAAZFAASGR-LFAAZFAASGR-GFAAZFAASGR-
GGFAAGNFSRR-VVAAGNFVSGR-LPGFAAZFAASGR-NFAAZFAASGR-VVAAGNFVSGR -VVAAGNFVSGR
1418.813 -HPLVAELPGYGR-HPLVAELPGGYR-RYVVPPVSXXR-HPLVAPNZGXXR-AGPLVAELPGXXR-MGPLVAELPGXXR-RGANTDCVXXR-RFDVPPVSXXR-RFVDPPVSXXR-
AGPLVAPNZGXXR-HPLVAPPVSXXR-MTNVAELPGXXR-RGANTFZVXXR-MGPLVAPNZGXXR-HEYVVPPVSXXR-ZPLVAELPGXXR-EPLVAELPGXXR-
HESNENVSXXR-RFVDELPGXXR-FYVVPPVSXXR-PZLVAELPGXXR
1554.924 -AALMPMRSSALPLR-ASXXSPLASSLNLR -ASXXLALPLSSALR-AANTWPLASSLNLR-MMSLALPLSSALR-MMSSPLPLSSALR-ASPFNSPLSSPXXK-
ASDENSPLSSPLAK-ASDENSPLSSPALK-ASDENSPLSSPSPK-ASDENSPLSSPPSK-MMSLSPPLSSALR-ANMSLALPLSSALR-AALMPMRPLSSALR-
GNMSLALPLSSALR-LMSLALPLSSALR-WMSLALPLSSALR-LMSLSPPLSSALR-CMSLALPLSSALR-ANMSSPLPLSSALR-BGPAAMSLALPLSSALR-
BPGAAMSLALPLSSALR-GNMSSPLPLSSALR-LMSSPLPLSSALR
1609.774 -MZGXXPSDYLFER-GGRGXXSSDYLFER-MZGXXSSDYLFER-GGZGXXSSDYLFER-GGZGXXPSDYLFER-AEGXXSSDYLFER-NZGXXSSDYLFER-SEGXXPSDYLFER-
NRGXXSSDYLFER-AEGXXPSDYLFER-GGRGXXPSDYLFER-NZGXXPSDYLFER-EEGXXSSDYLFER-MEGXXSSDYLFER-EEGXXPSDYLFER-LEGXXSSDYLFER-
TLZGXXSSDYLFER-SEGXXSSDYLFER-MEGXXPSDYLFER-LEGXXPSDYLFER
1666.797 -PLAGGGVDPGFPGPDPR-LPAGGVFFLYGPDPR-PLAGGVFFLYGPDPR-YVGGGVDPGFPGPDPR-LVGGGVDPGFPGPDPR-YVLTGLGVGWGPDPR-
LPAGGGVDPGFPGPDPR-MAGGGVDPGFPGPDPR-ATEGGVDPGFPGPDPR-LVGGVFFLYGPDPR-EEGGVDPGFPGPDPR-STEGGVDPGFPGPDPR-
STEGTGAVVGWGPDPR-NAGGGVDPGFPGPDPR-LEGGVDPGFPGPDPR-YVLTGAVVGWGPDPR-FGGVDPGFPGPDPR-EAGGTZVVGWGPDPR-
YVGGGVNGVGWGPDPR-EZGTGAVVGWGPDPR
1736.907 -BSGDGVNYTDMWPPER-GGVGVNYTDMWPPER-GRGVNYTDMWPPER-PDGVNYTDMWPPER-AADGVNYTDMWPPER-LVGVNYTDMWPPER-
NVGVNYTDMWPPER-VDGVNYTDMWPPER-GGVRNYTDMWPPER-DVGVNYTDMWPPER-NEPPWLPHYPVTAR-DEPPWLPHYPVTAR-GGVGVPWMXXYPVTAR-
PZDPWLPHYPVTAR-PFPPWLPHYPVTAR-GGVRPXXLPHYPVTAR-GGVGVPXXLPHYPVTAR-NEXXWLPHYPVTAR-TGEPPWLPHYPVTAR-VZDPWLPHYPVTAR
1793.950 -BGVELSGGYHPLEPPGER-BGVELSGGYHPLEPGPER-BGVELSGGYHPLEHNER-BGVELSGGYHPLENHER-BGPMLSGGYHPLEHNER-BGPMLSGGYHPLENHER-
BGPMLSGGYHPLEPXXER-BGVELSNYHPLEPXXER-BGVEEZGYHPLEPXXER-BGVEEAGGYHPLEXXER-BGVEEAGGYHPLEPXXER-BGPMEZGYHPLEPXXER-
BGPMLSNYHPLEPXXER-BGPMEAGGYHPLEXXER-BGDLLSGGYHPLEPXXER-BGPMEAGGYHPLEPXXER-BGVELSGGYHPLEPWPR-BGVELSGGYHPLEPPWR-
BGVZNSGGYHPLEPXXER-GPMLSGGYHPLEPXXR-BGVEEANYHPLEPXXER-BGPMEANYHPLEPXXER-BNLLSGGYHPLEPXXER-BGDLEAGGYHPLEPXXER
1850.950 -PFLVPANPGHYGGVDPR-PFLVPANPGHYGGVPDR-PFLVPANPGHYGGVVLR-PFLVPANPGHYGGVLVR-PFLVPANPGHYGRXXR-PFLVPANPGYHGGVXXR-
LELVPANPGHYGGVXXR-PETGAPANPGYHGGVXXR -DELVPANPGHYGGVXXR-PETZPANPGYHGGVXXR-PFLVPANPHYGGGVXXR-PFLVPANPHGYGGVXXR-
NELVPANPGHYGGVXXR-PFLVPANPGYHGRXXR-LMLVPANPGHYGGVXXR-PFLVPANPGHYNVXXR-LELVPANPGHYGRXXR-LRZGPANPGYHGGVXXR-
LELVPANPGYHGGVXXR-DELVPANPGHYGRXXR-PFLVPANPHYGGRXXR-LRZGPANPGHYGGVXXR-NRZGPANPGYHGGVXXR
1938.893 -BXXSGZFALGEYEEYVTK-GSZFALGEYEEYVTK-SGZFALGEYEEYVTK-GSZFALGECCTYVTK-GSZFALGECCTYVTK-GSZFALGEYEEYVTK-
NMGGAFALGEYEEYVTK-NTSGAFALGECCTYVTK-YZEEYEGLVTSGSMK-LMGGAFALGEYEEYVTK-SGGAFALGEYEEYVTK-SGZFALGEYEEYVTK-
GSZFALGEYEEYAEK-GSGAFALGECCTYVTK-LTSGAFALGECCTYVTK-GSZFALGEYEEYVTK-TVVGAFALGEYEEYVTK-GSZFALGEYEEYVTK-
YNNVLGEYSAFVCGK-NTSGAFALGEYEEYVTK
2003.131 -SGLLHWSLAAGELVVVHR-SGLLHWSALAGELVVVHR-SGLLHWSSPAGELVVVHR-SGLLHWSPSAGELVVVHR-DVDLGACLAZELGZYGGK-
DVDLGACLAZELGZGYGK-PVDLGACLAZELGZXXGK-NVDLGACLAZELGZXXGK-RGDLGACLAZELGZXXGK-NGSALGACLAZELGZXXGK-DVSVVACLAZELGZXXGK-
DVSDGZELAZELGZXXGK-DVSDGZEXXAGELVVVHR-LVDLGACLAZELGZXXGK-NFGSLPDLAZELVSXXK-MAVDGZEXXAGELVVVHR-PGSALGACLAZELGZXXGK-
PVSVVACLAZELGZXXGK-DSVDGZEXXAGELVVVHR-PVSDGZELAZELGZXXGK-RGSDGZEXXAGELVVVHR-NVSDGZEXXAGELVVVHR-LGSALGACLAZELGZXXGK-
SRGSLPDLAZELVSXXK

Figure 3. The MS BLAST sequence query string composed from the candidate sequences determined by PredictSequence. Partial sequences that were subsequently matched to various OATs by MS BLAST are highlighted.

above). We therefore reasoned that a number of unique peptide sequences in the MS BLAST query would only slightly exceed the number of fragmented peptides and performed the simulation experiments with the queries comprising 10−50 unique nonredundant peptide sequences. The simulating program developed in-house fetched protein sequences from the strongly nondegenerate database of globular proteins SCOP40.[42] Reversed peptides of 11 amino acid residues (the average length of tryptic peptides) were assembled into query strings and submitted for MS BLAST searching using the same settings and the same database as described above. For each top hit protein, the total score and the number and scores of individual HSPs were recorded. Figure 4 presents the simulated distribution of the total scores of the top protein hits matching queries consisting of 50 nonredundant random peptide sequences. The "steps" at the plot correspond to various numbers of reported HSPs. Separate conditional distributions for the fixed numbers of reported HSPs (as shown in the inset) were used to calculate the thresholds that are presented in Table 2. Figure 4 demonstrates that the total score as such is not a reliable criterion of the statistical significance of the match. For example, the total score of 111 represents 27% probability that such hit occurred at random (Figure 4). However, if the same score of 111 was observed when exactly two HSPs were matched, the probability that this hit is a false positive is lower than 1% (Figure 4, inset).

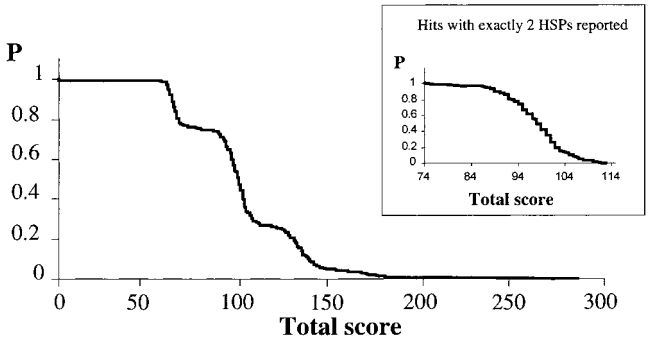

Figure 4. Simulated distribution of total scores of the top hits of MS BLAST searching with the queries composed of 50 random peptide sequences. *Y*-Axis presents the estimate of the probability (*P*) that the total score of the top hit exceeds the given value (plotted at *X*-axis). The inset presents conditional probability distribution of the total score given that exactly two HSPs were reported for top hits.

How to evaluate the significance of MS BLAST hits using the thresholds from Table 2? First compare the score of the top-ranked HSP of the hit with the threshold score expected for a single random matching HSP calculated for a given number of fragmented peptide precursors (Table 2). If the score of the HSP of the hit is higher than the threshold, the protein is positively identified. If the score is below or very close to the threshold, then the score of the second-ranked HSP is added and the sum is compared with the threshold expected for two randomly matching HSP (Table 2). Again, if the sum is higher than the threshold, the sample is positively identified. If not, the procedure

(42) Lo Conte, L.; Ailey, B.; Hubbard, T. J.; Brenner, S. E.; Murzin, A. G.; Chothia, C. *Nucleic Acids Res.* **2000**, *28*, 257−259.

Table 2. Approximate Threshold Scores Determined via Computer Simulation Experiments for Queries Composed of Random Peptide Sequences[a]

no. of reported HSPs	no. of unique peptides in the query		
	10	20	50
1	68	72	75
2	102	106	111
3	143	146	153
4	177	<208[c]	<180[c]
5	<238[c]	n.o.[b]	<212[c]
6	n.o.	n.o.	<275[c]
7	n.o.	n.o.	<285[c]

[a] A total of 99% of top hits matching the specified number of HSPs had the score below the presented thresholds. Approximately 1000 MS BLAST searches were performed for each experiment. [b] n.o., not observed. No random hits with specified number of HSPs were observed. [c] The calculated value is statistically unreliable because just a few hits matching with the specified number of HSPs were observed. In those cases, the maximal score from the ones observed is presented solely as a reference.

is repeated with adding the third-ranked HSP and so forth. It is important to start the comparison from the top-scoring HSP of the protein hit since otherwise the low-scoring HSPs may erroneously decrease the statistical significance of the match.

Let us consider the identification of the canine homologue of OAT as an example (Table 1). Since tandem mass spectra of 15 peptide precursor ions were acquired we assumed that the query (Figure 3) contains less than 20 of unique nondegenerate peptide sequences. The score of the top HSP (peptide sequence LGEYEEYVTK) is 71 (Table 1) and is very close to the threshold of statistical significance for a single matched HSP that is 72 (Table 2). We then add the score of the second-ranked HSP (sequence EEGXXSSDYLFER) that is 69 and compare the sum with the threshold for two matched HSPs that is equal 106. Now the score of the sum (140) significantly exceeds the threshold and positive identification can be claimed.

Thus, Table 2 makes possible evaluation of the statistical significance of hits independently of scores of randomly matching proteins or, in other words, independently of "information noise". Proteins can be positively identified even if only a single peptide was matched with the score that is below the highest total score of random matches.

However, caution has to be observed when using the calculated thresholds that are provided here merely for guidance rather than for use as the absolute criteria. It is always prudent to take into account indirect evidence that may (or may not) verify the match, such as the presence of conserved sequence motifs, similarity between proteins of the same family, etc.

All examples of identified proteins listed in Table 1 and Table 3 were significant with respect to the thresholds listed in Table 2, while all top false positive hits were not significant.

Identification of Proteins from *P. pastoris*. We tested the proposed strategy in a small-scale proteomic project involving identification of proteins purified from the methylotrophic yeast *P. pastoris*. Although *P. pastoris* has been widely used as a host organism for protein expression,[43] its genome is unknown and we are not aware of any genomic sequencing project currently underway. By now the combined Swissprot and Swissnew databases contain 22 complete sequences of *P. pastoris* proteins, which share 35−85% of sequence identity with corresponding proteins from various fungi. We note the important fact that *P. pastoris* is only distantly related to the much explored fungi organism *S. cerevisiae*, the genome of which has been publicly available since 1996,[27,44] so this investigation is a fairly stringent test of the proposed technique.

The statistics of protein identification are presented in Table 3, showing that 15 bands out of 19 attempted were successfully identified.

On average, 25 peptide masses per protein were included in the searching lists, but peptide mass fingerprinting identified only 2 out of a total of 15 proteins, because of the low similarity of *P. pastoris* proteins to known proteins. Although less than half of the full length sequence of *P. pastoris* elongation factor 3 (p120) was available in a database, better than 6 ppm mass accuracy allowed unambiguous identification of the protein. The protein p25 turned out to be a member of a conserved family of 14-3-3 proteins. Cross-species mass fingerprinting identified its homologue, and the match was confirmed by MASCOT searching.

Identifications produced by MASCOT and MS BLAST were in good agreement with the exception of sample p50. MASCOT hit a single peptide from a bacteriophage protein, and the hit was at the verge of statistical significance. MS BLAST identified a different protein (a pyruvate kinase, MW 55 100), making it a likely homologue of the protein to be identified. The bacteriophage protein was not even included in the list of MS BLAST hits. Furthermore, the mass of the bacteriophage protein is 14 kDa, and it does not have apparent homologues in fungi. Thus, we concluded that in this case the MASCOT hit was false.

On average MASCOT confidently matched one to two peptides out of 10 uploaded MS/MS spectra, whereas five peptides were assigned via MS BLAST (Table 3), thus increasing confidence in the MASCOT hits. Importantly, MS BLAST identified four proteins from the samples in which both mass fingerprinting and MASCOT searching failed (Table 3 and Table 4). A full list of MS BLAST peptide sequence alignments is provided as Supporting Information.

Although the technique could identify proteins in mixtures (samples p65 and p17), in no case were all fragmented peptide ions assigned to protein hit(s). Therefore, it is prudent to note that yet other unidentified proteins might still be present in the sample. Such a possibility can never be ruled out unless the full length sequence of the protein is obtained and screened against the acquired MS/MS spectra.

CONCLUSION AND PERSPECTIVES

The present paper demonstrates that characterization of the proteome of organisms with unknown genomes can be carried out by a combination of peptide mass fingerprinting, MS/MS-based database searching, and MS BLAST protocol. In a single experiment, MALDI-QqTOF mass spectrometry provides data for these three database mining strategies. The combination of MALDI-QqTOF mass spectrometry and MS BLAST searching thus allows facile identification of proteins sharing only moderate

(43) Cereghino, J. L.; Cregg, J. M. *FEMS Microbiol. Rev.* **2000**, *24*, 45−66.

(44) Goffeau, A. *FEBS Lett.* **2000**, *480*, 37−41.

Table 3. Identification of Proteins from *P. pastoris*

		identified by MASCOT			identified by PredictSequence and MS BLAST			
band[a]	MS/MS spectra	top hit	org	peptides matched[b]	peptide sequences in the query	top hit	org	peptides matching top hit + other homologues
p17	10	AF202054 nucleoside diphosphate kinase-Z3	*DR*	1c	1989[c]	Q13232 nucleoside diphosphate kinase 3	*HS*	3 + 1[d]
						P22011 peptidyl-prolyl cis−trans isomerase	*CA*	3 + 0[d]
p18	9	n.i.[e]			197	n.i.		
p20	13	n.i.			291	Q11118 WOS2 protein	*SP*	4 + 1
p24	7	AF149421 thiol-specific antioxidant-like protein	*CA*	1c	112	O74887 thioredoxin peroxidase	*SP*	3 + 1
p25	8	proteins of 14-3-3 family, various species		identified by mass fingerprinting, 8 peptides matched at 10 ppm tolerance [f]				
p27	9	P00942 triosephosphate isomerase TPI1	*SC*	2c	160	P04828 triosephosphate isomerase	*EN*	3 + 1
p28	10	n.i.			201	n.i.		
p40		P30575 enolase 1	*CA*	1c	168	P30575 enolase 1	*CA*	4 + 3
p45	19	n.i.			423	n.i.		
p50	9	AE004507 hypothetical protein of bacteriophage Pf1	*PA*	1c	942	P52489 pyruvate kinase 2	*SC*	3 + 0
p52	13	U75310(1) pyruvate decarboxylase 1(2)	*PS*	2c	202	P26263 pyruvate decarboxylase 3	*SC*	6 + 0
p55	9	n.i.			192	P32527 zuotin	*SC*	4 + 0
p57	8	n.i.			178	n. i.		
p65	9	O94039 transketolase I	*CA*	1c	331	O94039 transketolase I	*CA*	4 + 1
		P12398 heat shock protein SSC1	*SC*	1c				
p70	14	protein SSB1(2)	*SC*	3c	263	P41770 heat shock protein SSB	*KM*	9 + 3
p75	11	AF111194 heat shock protein 70	*PB*	1c	234	P41887 heat shock protein 90	*SP*	5 + 3
		M26044 hsc82 protein	*SC*	2c				
p80	19	proteins of HSP90 family, various species		5c + 2nc	423	P02829 heat shock protein HSP82	*SC*	14 + 1
p95	10	AF107287 elongation factor 2	*CG*	2c + 2nc	217	AAG09782 elongation factor 2	*FN*	7 + 1
p120		O93813 elongation factor 3	*PP*	identified by mass fingerprinting; 17 peptides matched at 10 ppm tolerance				

[a] Bands excised from a polyacrylamide gel are named by their apparent molecular weights. [b] c, peptides matched with statistically confident score; nc, peptides matched to the same protein with lower score. [c] Sequences predicted by PredictSequence (BioMultiview) (171) and BioAnalyst (1818) were combined. [d] pp8 is a mixture of two proteins. Sequences matching Q13232 or P22011 originated from different peptide precursors. [e] n.i., not identified. [f] Confirmed by MASCOT. Three peptides matched with statistically significant scores.

Table 4. Peptide Sequence Alignments for Proteins Identified by MS BLAST Only[a]

Band	Best hit	High Scoring Pairs
p17	P22011	Query: BTLENFR PZFTLZGGDFDN HVVFGEVVDGLDVVZ +| |||| |+| |+||| | |||||| |||\+|+ Sbjct: 29 KTAENFR 35 56 PQFMLQGGDFTN 67 124 HVVFGEVTDGLDIVK 138
p20	Q11118	Query: PEVLWAZR LEFFEDVDVEK EEFWPR RLXXLHTDFDR |||||+| |+|| |+||| ||||| || | |||| Sbjct: 8 PEVLWAQR 15 63 IDFFKDIDVEK 73 94 EEFWPR 99 106 RLHWLRTDFDR 116
p50	P52489	Query: EFHQ VFASFLR DNFDELLEVTD |||| |||||| ||||||||| Sbjct: 60 EFHQ 63 212 VFASFIR 218 250 DNFDEILEVTD 260
p55	P32527	Query: ADLYA YDFFEAWGPLFXXEGR DSWR DEDVPDDTSNR ||||| ||\+|||| | | | |||| |||||| ||| Sbjct: 96 ADLYA 100 182 YDFYEAWGPVFEAEAR 197 228 DSWR 231 237 DEDVPDDSSNR 247

[a] Full list of sequence alignments is provided as Supporting Information.

sequence similarity with their known homologues. Together with the rapid growth of databases, the technology described here may therefore eliminate the major obstacle in expanding the scope of proteomic research well beyond organisms "blessed" by genomic sequencing.

A rapidly switchable ESI/MALDI ion source[36,45] offers another intriguing perspective for improving the quality of data. MALDI and ESI produce different patterns of peptides when the same

(45) Krutchinsky, A. N.; Loboda, A. V.; Spicer, V. L.; Dworschak, R.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 508−512.

protein digest is analyzed,[46] so tandem mass spectra acquired in MALDI and ESI modes on the same quadrupole time-of-flight instrument could be combined and used in a single packet for MASCOT and MS BLAST searching.

Although MS BLAST helps one to tolerate a very high level of "information noise", it is crucially important whether or not the spectra interpretation software is able to call a correct stretch of peptide sequence for each MS/MS spectrum. This emphasizes the urgent need for further development of software for automated de novo sequencing—software that is specifically tailored for particular ionization and fragmentation techniques.[47−48]

Abbreviated names of organisms: *CA*, *Candida albicans*; *CE*, *Caenorhabditis elegans*; *CG*, *Candida glabrata*; *DA*, *Drosophila ananassae*; *DM*, *Drosophila melanogaster*; *DR*, *Danio rerio*; *EN*, *Emericella nidulans*; *FN*, *Filobasidiella neoformans*; *HS*, *Homo sapiens*; *KM*, *Kluyveromyces marxianus*; *MM*, *Mus musculus*; *PA*, *Pseudomonas aeruginosa*; *PP*, *Pichia pastoris*; *PB*, *Paracoccidioides brasiliensis*; *PS*, *Pichia stipitis*; *SC*, *Saccharomyces cerevisiae*; *SP*, *Schizosaccharomyces pombe*.

(46) Shevchenko, A.; Loboda, A.; Ens, W.; Schraven, B.; Standing, K. G.; Shevchenko, A. *Electrophoresis*, in press.

(47) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327−342.

(48) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10313−10317.

ACKNOWLEDGMENT

The work at Manitoba was supported by grants from the Natural Sciences and Engineering Research Council of Canada, from MDS Sciex, and from the U.S. National Institutes of Health (GM59240). We acknowledge the excellent technical assistance of V. Spicer, J. McNabb, and S. Jilkine (University of Manitoba). We are grateful for Drs. J. Fullerkrug and K. Simons (MPI CBG, Dresden) for providing a sample of canine proteins; Drs. F. Severin, A. Pozdniakovsky, and A. Hyman (MPI CBG, Dresden) and Dr. S. Weeds (EMBL, Heidelberg) for supporting experiments with *P. pastoris*; and Dr. Yan P. Yuan (EMBL) for his valuable assistance with the Web interface of WU-BLAST.

SUPPORTING INFORMATION AVAILABLE

A full list of MS BLAST peptide sequence alignments. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review November 21, 2000. Accepted March 6, 2001.

AC0013709