*Anal. Chem.* **2002,** *74,* 2529−2534

# Mass Spectrometry-Based Proteolytic Mapping for Rapid Virus Identification

Zhong-Ping Yao, Plamen A. Demirev, and Catherine Fenselau*

*Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742*

**A novel method is proposed for rapid identification of viruses and other organisms that show a low number of biomarkers, based on the construction of databases of organism-specific tryptic peptide masses. The peptide products of any protease that cuts at specific residues can be accommodated. Experimentally, a sample of intact virus, e.g., one collected from the atmosphere, is digested with a selective protease for a short time, and the digestion products are analyzed by MALDI-TOF mass spectrometry without fractionation or purification. In the present proof of concept, the Sindbis virus AR 339 was identified by using the masses of observed tryptic peptide products to query a database composed of tryptic peptide masses generated *in silico* for six viruses whose genomes have been sequenced. Two algorithms were tested for identification—a direct score-ranking algorithm and an algorithm that evaluates the probability of random matching. The Sindbis virus was unambiguously identified by either approach. The influence of factors such as experimental mass accuracy, number of missed cleavages, and database size on the identification algorithms has also been evaluated, with the objective of extending the approach to other microorganisms.**

Various mass spectrometric approaches have been vigorously applied in the last several years to characterize proteins in known viruses and also to characterize unknown viruses. In the former category, studies have been reported of protein mutations,[1,2] genetically engineered proteins,[3] capsid protein dynamics,[4,5] and surface glycoprotein conformations.[6] In most of these, proteolysis was carried out on the intact virus, often with incubation time as a variable. In addition to trypsin, pepsin[6] and thermolysin[7] have been used. Of particular relevance to the present work, both proteases[6] and glycosidases[8] have been successfully used on intact Sindbis virus. In the second category, identification of unknown viruses by mass spectrometry, preliminary efforts have been reported to weigh intact viruses using electrospray ionization.[9−11] In addition, there is considerable interest in identifying viruses rapidly in portable MALDI-TOF instruments. This approach is usually based on matching experimentally observed biomarker masses with a database of protein masses (including those predicted from sequenced genomes).[12−15] The answer to the question, "What is this microorganism?", requires a different strategy than identification of an unknown protein, and at least one new bioinformatics algorithm is being developed to address this.[16−18]

Peptide mass maps obtained by proteolytic digestion are widely used to search databases to identify purified proteins.[19,20] Some advances have also been made in the extension of this approach to small mixtures of proteins,[21] although sequence information obtained by tandem mass spectrometry experiments is more widely used with mixtures.[22]

In the present study, we propose a new bioinformatics strategy in which a database is assembled to contain the masses of all the peptides predicted from residue-specific protease cleavage (e.g., by trypsin) of all the proteins in each microorganism. In this

* To whom correspondence should be addressed. Phone: 1-301-405-8614. Fax: 1-301-405-8615. E-mail: fenselau@wam.umd.edu.

(1) Lewis, J. K.; Bendahmane, M.; Smith, T. J.; Beachy, R. N.; Siuzdak, G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 8596−8601.
(2) She, Y. M.; Haber, S.; Seifers, D. L.; Loboda, A.; Chernushevich, I.; Perreault, H.; Ens, W.; Standing, K. *J. Biol. Chem.* **2001**, *276*, 20039−20047.
(3) Carrion, M.; Smith, J.; Harris, B.; McVey, D. *Proceedings of the 48th ASMS Conference on Mass Spectrometry and Allied Topics*; Long Beach, CA, June 2000; pp 1205−1206.
(4) Lewis, J. K.; Bothner, B.; Smith, T. J.; Siuzdak, G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6774−6778.
(5) Broo, K.; Wei, J.; Marshall, D.; Brown, F.; Smith, T. J.; Johnson, J. E.; Schneemann, A.; Siuzdak, G. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2274−2277.
(6) Phinney, B. S.; Blackburn, K.; Brown, D. T. *J. Virol.* **2000**, *74*, 5667−5678.
(7) Bark, S.; Muster, N.; Yates, J.; Siuzdak, G. *J. Am. Chem. Soc.* **2001**, *123*, 1774−1775.
(8) Kim, Y. J.; Freas, A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 1544−1548.
(9) Chernushevich, I.; Ens, W.; Standing, K. In *New Methods for the Study of Biomolecular Complexes*; Ens W., Standing K., Chernushevich, I., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1998; pp 101−116.
(10) Tito, M. A.; Tars, K.; Valegard, K.; Hajdu, J.; Robinson, C. V. *J. Am. Chem. Soc.* **2000**, *122*, 3550−3551.
(11) Fuerstenau, S. D.; Benner, W. H.; Thomas, J. J.; Brugidou, C.; Bothner, B.; Siuzdak, G. *Angew. Chem., Int. Ed.* **2001**, *40*, 541−544. Corrigendum: *Angew. Chem., Int. Ed.* **2001**, *40*, 982.
(12) Thomas, J. J.; Falk, B.; Fenselau, C.; Jackman, J.; Ezzell, J. *Anal. Chem.* **1998**, *70*, 3863−3867.
(13) Birmingham, J.; Demirev, P.; Ho, Y. P.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 604−606.
(14) Tan, S. W.-L.; Wong, S.-M.; Kini, R. M. *J. Virol. Methods* **2000**, *85*, 93−99.
(15) Bundy, J. L.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 751−757.
(16) Pineda, F.; Lin, J.; Fenselau, C.; Demirev, P. *Anal. Chem.* **2000**, *72*, 3739−3744.
(17) Demirev, P.; Pineda, F.; Lin, J.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 4566−4573.
(18) http//: www.infobacter.jhuapl.edu.
(19) Henzel, W. J., Billeci, T. M., Stults, J. T.; Wong, S. C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.
(20) Shevchenko, A.; Jensen, O.; Podtelejnikov, A.; 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.
(21) ProFound: http://129.85.19.192/profound_bin/WebProFound.exe.
(22) Wolters, D.; Washburn, M.; Yates, J. *Anal. Chem.* **2001**, *73*, 5683−5690.

10.1021/ac0200217 CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/30/2002

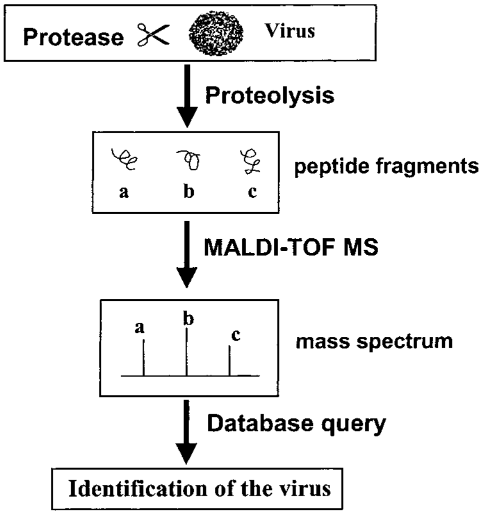

Figure 1. Strategy for rapid virus identification by partial proteolysis, mass spectrometry, and database queries.

strategy, the experimentally observed peptide masses are searched against the database to identify the virus directly (Figure 1). Proteins are not identified, nor are their masses used in the search.

Sindbis virus AR339 (SIN) is used as the model analyte to test the method. Its genome is completely sequenced.[23,24] The sequenced genomes of the bacteriophage MS2, Ross River virus strain NB5092, tobacco mosaic virus U2, human adenovirus strain 5, and Lelystad virus were processed to create a model peptide database for searching. Each peptide entry in the database contains information about the organism from which it originates. A score based on the number of matches between masses of experimentally observed peptides and peptides in the database, within a specified mass tolerance window, was initially used for virus identification. In this manner, we demonstrate that rapid identification can be achieved by proteolytic mapping and database query. Further, we evaluate a statistical approach, previously introduced for microorganism identification using intact protein biomarkers,[16] that calculates the probability of random matches. Such an approach is found to improve the specificity of viral identification.

## EXPERIMENTAL SECTION

**Materials.** Sindbis virus strain AR339 was purchased from American Type Culture Collection (Manassas, VA) and propagated and purified as previously described.[8] Sequencing grade modified trypsin and trypsin resuspension buffer (50 mM acetic acid) were provided by Promega (Madison, WI). The matrix—α-cyano-4-hydroxycinnamic acid (97%)—and trifluoroacetic acid (TFA; 99%) were obtained from Aldrich Chemical Co. (Milwaukee, WI). Ammonium bicarbonate and acetone were from J. T. Baker Chemical Co. (Phillipsburg, NJ). 2-Propanol (HPLC grade) was from Fisher Scientific (Fair Lawn, NJ) and 2-mercaptoethanol from Sigma Chemical Co. (St. Louis, MO). Nitrocellulose was purchased from Bio-Rad (Hercules, CA). CAUTION: Sindbis virus is classified as a "Biohazard level two" (BL2) microorganism, and proper handling procedures should be followed.[25]

**Digestion of the Virus.** Sindbis virus suspension (10 $\mu$L) was mixed at room temperature with 10 $\mu$L of 100 mM ammonium bicarbonate buffer (pH 8) containing 2-mercaptoethanol (2%). Then 2 $\mu$L of trypsin (0.1 $\mu$g/$\mu$L in 50 mM acetic acid) was added. At various digestion times, 0.2 $\mu$L of digestion solution was transferred for MALDI-MS analysis.

**Mass Spectrometry.** MALDI-TOF analysis was performed on a Kompact MALDI 4 mass spectrometer (Kratos Analytical Instruments, Chestnut Ridge, NJ), equipped with a nitrogen laser (337 nm). Positive ion mass spectra were acquired in the linear mode with delayed extraction and at an accelerating voltage of ~20 kV. The thin-layer sample preparation technique[26] was used. Initially, a thin matrix layer was formed by depositing 0.2 $\mu$L of a mixture containing 20 $\mu$g/$\mu$L α-cyano-4-hydroxycinnamic acid and 5 $\mu$g/$\mu$L nitrocellulose in acetone/2-propanol (1:1). A droplet of 0.2 $\mu$L of 1% TFA was added to the thin layer prior to loading the sample. Typically 80 single laser shot traces were summed for each spectrum. The instrument was internally calibrated with known peaks corresponding to autolysis peptides of trypsin.

**Database Construction and Query.** All proteins contained in NCBI proteome databases[27] and belonging to six different viruses, namely, Lelystad virus (LV), Sindbis virus (SIN) strain AR339, Ross River virus (RRV) strain NB5092, human adenovirus type 5 (AD5), tobacco mosaic virus (TMV) U2, and bacteriophage MS2 (MS2), were used to construct the database. The proteins were digested *in silico* with trypsin using the PeptideMass software.[28] Up to four missed cleavage sites (#MC) were allowed for each *in silico*-digested protein. All proteolytic peptides (their sequences, masses, and number of missed cleavages) for each microorganism were uploaded to a PC for further statistical analysis. The peptide database has been restricted within a mass range from 1000 ($m_{\min}$) to 4000 Da ($m_{\max}$). The number of tryptic peptide fragments from all proteins of the virus observed in the MALDI mass spectrum was denoted by $K$. All experimentally observed masses were compared with the masses of the *in silico*-generated tryptic peptides from each virus. A match was counted if the experimentally observed mass coincided with the mass of a database peptide, within a selected $\Delta m$—mass accuracy (tolerance) window. A score ($S$) for each experimentally observed spectrum was calculated according to $S = k_{\text{match}}/K$, where $k_{\text{match}}$ is the number of matched peptides for each virus in the database. The virus that yielded the highest score was identified as the target organism. In addition, significance level testing, a statistical approach previously introduced for microorganism identification using intact protein biomarkers,[16] was applied to calculate the probability of random matching (vide infra) on a scale of 0 to 1.

## RESULTS AND DISCUSSION

**Mass Spectrometry.** Digestion of the SIN virus for 2 min results in the observation of more than 20 tryptic peptides in a MALDI spectrum (Figure 2a). A similar number of peptide peaks is observed in the spectrum after digestion for 5 min (Figure 2b),

[23] Strauss, E. G.; Rice, C. M.; Strauss, J. H. *Virology* **1984**, *133*, 92−110.

[24] McKnight, K. L.; Simpson, D. A.; Lin, S.-C.; Knott, T. A.; Polo, J. M.; Pence, D. F.; Johannsen, D. B.; Heidner, H. W.; Davis, N. L.; Johnston, R. E. *J. Virol.* **1996**, *70*, 1981−1989.

[25] Centers for Disease Control and Prevention/National Institutes of Health. *Biosafety in Microbiological and Biomedical Laboratories*, 4th ed.; U.S. Government Printing Office: Washington, DC, 1999.

[26] Kussmann, M.; Roepstorff, P. In *Protein and Peptide Analysis: New Mass Spectrometric Applications*; Chapman, J. R., Ed.; Humana: Totowa, NJ, 2001; pp 405−424.

[27] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[28] PeptideMass: http://ca.expasy.org/tools/peptide-mass.html.

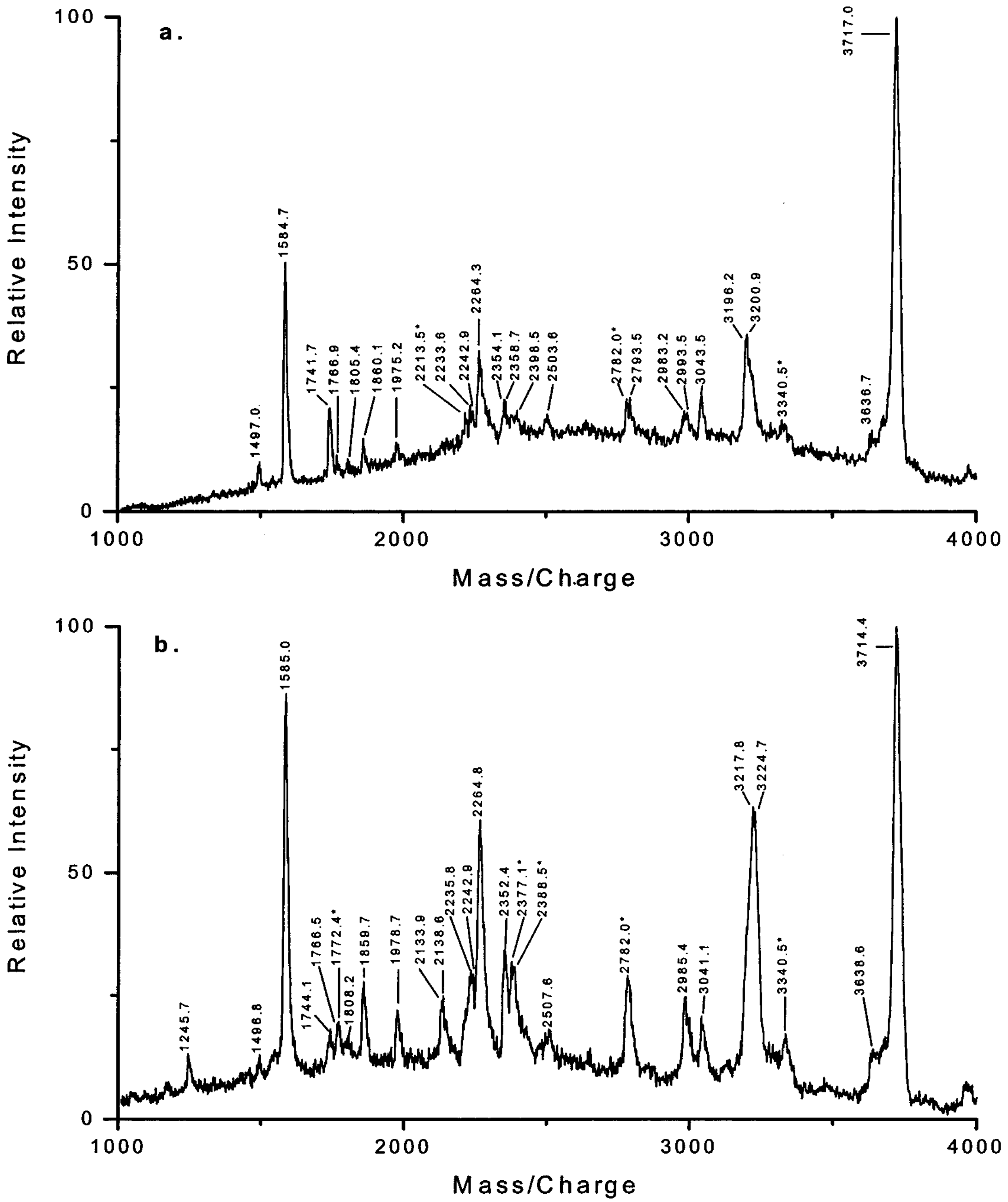

Figure 2. Positive ion MALDI-TOF mass spectra of tryptic products of Sindbis virus AR339 after (a) 2-min digestion with trypsin and (b) 5-min digestion with trypsin. Autolysis peaks of trypsin are marked with asterisks.

although some of the peptides are different. Digestion times longer than 5 min did not increase the number of observed tryptic peptides. This short time scale is required for rapid analysis in the battlefield. The Sindbis virus contains seven proteins: the capsid protein, membrane glycoproteins E1 and E2, and four nonstructural proteins nsp1, nsp2, nsp3, and nsp4.[23,29] The possible origin of each observed tryptic peptide is obtained by comparing its mass with the masses of all the *in silico*-generated tryptic peptides of SIN proteins. Peptides from both structural and nonstructural proteins are observed (Table 1), suggesting that at least some viruses were disrupted during processing. Under the present experimental conditions, fewer than 25 tryptic peptides from the Sindbis virus are experimentally observed, compared to more than 1000 theoretically generated peptides in the mass range 1000−4000 Da. Although abundant ion signals below $m/z$ 1000 were also present, we did not consider these since not all could be correlated to tryptic peptides. For instance, strong signals between $m/z$ 730 and 780 are thought to originate from intact phosopholipids.[30] A few species above $m/z$ 1000 cannot be successfully matched as originating from translated proteins and may represent glycolipids or other secondary metabolites.

[29] Schlesinger, S.; Schlesinger, M. J. In *Fundamental Virology*, 3rd ed.; Fields, B. N., Knipe, P. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizman, B., Straus, S. E., Eds.; Lippincott-Raven: Philadelphia, 1996; Chapter 17, pp 523−539.

[30] Bryant, D. K.; Orlando, R. C.; Fenselau, C. *Anal. Chem.* **1991**, *63*, 1110−1114.

Table 1. Masses[a] of Tryptic Peptides, Observed in MALDI Spectra at Various Digestion Times, and Their Tentative Assignment

| 2 min | | 5 min | |
|---|---|---|---|
| obsd mass | possible origin[b] | obsd mass | possible origin[b] |
| 1497.0 | nsp2 | 1245.7 | capsid, E2 |
| 1584.7 | capsid, nsp2 | 1496.8 | nsp2 |
| 1741.7 | capsid, E1, E2 | 1585.0 | capsid, nsp2 |
| 1766.9 | nsp3 | 1744.1 | E2 |
| 1805.4 | E2, nsp3 | 1766.5 | nsp3 |
| 1860.1 | capsid, E1, nsp1 | 1808.2 | E2 |
| 1975.2 | nsp2, nsp3 | 1859.7 | capsid, E1, nsp1 |
| 2233.6 | E2, nsp1, nsp2 | 1978.7 | E2, nsp4 |
| 2242.9 | capsid, E2, nsp2, nsp3, nsp4 | 2133.9 | E2 |
| 2264.3 | E2, nsp1 | 2138.6 | E2 |
| 2354.1 | E2, nsp3, nsp4 | 2235.8 | capsid, E2, nsp2, nsp3 |
| 2358.7 | E2, nsp3 | 2242.9 | capsid, E2, nsp2, nsp3, nsp4 |
| 2398.5 | nsp1, nsp4 | 2264.8 | E2, nsp1 |
| 2503.6 | N/A[c] | 2352.4 | E1, E2, nsp3, nsp4 |
| 2793.5 | N/A[c] | 2507.6 | capsid |
| 2983.2 | nsp2 | 2985.4 | E2, nsp2, nsp3 |
| 2993.5 | E2 | 3041.1 | nsp2, nsp3 |
| 3043.5 | E2 | 3217.8 | capsid |
| 3196.2 | nsp1, nsp4 | 3224.7 | E2, nsp2, nsp3 |
| 3200.9 | E2, nsp3 | 3638.6 | capsid, E2, nsp3 |
| 3636.7 | capsid, E2, nsp3 | 3714.4 | E1, nsp2, nsp4 |
| 3717.0 | capsid, nsp2, nsp4 | | |

[a] Mass range 1000−4000 Da, trypsin autolysis products not included. [b] Obtained by comparison with *in silico* tryptic digestion peaks of the Sindbis proteins with mass tolerance ±2 Da and maximum missed cleavage 4. [c] No matched Sindbis proteins.

Table 2. Viruses Used in Construction of the Database

| virus[a] | genome size (base pairs) | no. of different proteins | no. of tryptic peptides included |
|---|---|---|---|
| LV[35] | 15111 | 8 | 1033 |
| SIN[23,29] | 11703 | 7 | 1025 |
| RRV[36] | 11657 | 7 | 947 |
| AD5[37] | 35953 | 11 | 936 |
| TMV[38] | 6355 | 3 | 717 |
| MS2[39] | 3569 | 4 | 322 |

[a] References for the viral genome and protein sequences are given.

**Protein Database Searches.** Initially it seemed possible that the observed masses from the MALDI spectrum of the tryptic digest of intact Sindbis virus could be used to search protein databases, in an effort to identify individual proteins and, thereby, the microorganism. This strategy has the major drawback that protein identification search engines based on tryptic peptide mapping are designed to work with pure proteins or very small mixtures. Nonetheless, its feasibility to identify microorganisms was tested using three different on-line search engines containing the "virus" taxonomy classifier, namely, PeptIdent,[31] Mascot,[32,33] and ProFound.[21,34] We performed queries with all of them by using the SIN proteolytic peptide data (Table 1). No SIN protein was found among the first 20 viral protein candidates using Mascot. Search by PeptIdent resulted in a SIN protein hit ranked fifth only for the 5-min digestion data. The third engine tested—ProFound—has been recently upgraded to identify individual proteins in a mixture of up to four proteins.[21] A sindbis protein was ranked first for the 5-min digestion. However, no SIN proteins could be identified among the first 20 candidates of the search results for the 2-min digestion.

**Construction and Searches of a Tryptic Organism Database.** New strategies are evidently needed for a general method for microorganism identification with higher success rates, based on peptide mapping. Using the same experimental data (Table 1), we performed a query in the specially constructed database that contained ~5000 tryptic peptides from the six viral proteomes (Table 2).[35−39] For each spectrum, the score values for each of the six viruses was determined during the query (Table 3). SIN had the highest score value for all observed spectra at the four digestion times, even with a tolerance window of 2 Da (Table 3; data for 10 and 15 min not shown). For other organisms, random, coincident matches were observed. To evaluate statistically the probability of random matches in different species, significance testing was performed.

**Significance Testing.** Significance hypothesis testing has been used previously to estimate the likelihood for false matching of tryptic peptides for individual protein identification.[40] Significance testing algorithms have also been exploited to improve methods for microorganism identification by database queries using mass values of intact protein biomarkers.[16] In the latter case, a quantity—the significance level—α has been introduced, with values ranging between 0 (best) and 1 as a means to quantify statistically the probability for a random "hit". A lower α value corresponds to a higher confidence level of unambiguous identification. It has been demonstrated that the significance level is a function both of proteome density (i.e., the number of intact proteins per mass interval) and of mass accuracy.[16,17]

In the following, we follow closely the approach already developed in ref 16. Here we assume that tryptic peptides are uniformly distributed within the mass range from 1000 to 4000 Da. This assumption is clearly fulfilled for mass distributions of tryptic peptides with up to four missed cleavages (Figure 3a). In contrast, mass distributions (Figure 3b) for proteolytic peptides formed with no missed cleavages exhibit exponential behavior.[41]

[31] PeptIdent: http://ca.expasy.org/tools/peptident.html.

[32] Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.

[33] Mascot: http://matrixscience.com.

[34] Zhang, W. Z.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482−2489.

[35] Meulenberg, J. J. M.; Hulst, M. M.; De Meijer, E. J.; Moonen, P. L. J. M.; Den Besten, A.; De Kluyver, E. P.; Wensvoort, G.; Moormann, R. J. M. *Virology* **1993**, *192*, 62−72.

[36] Faragher, S. G.; Meek, A. D. J.; Rice, C. M.; Dalgarno, L. *Virology* **1988**, *163*, 509−526.

[37] (a) Chroboczek, J.; Bieber, F.; Jacrot, B. *Virology* **1992**, *186*, 280−285. (b) Shenk, T. In *Fundamental Virology*, 3rd ed.; Fields, B. N., Knipe, P. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizman, B., Straus, S. E., Eds.; Lippincott-Raven: Philadelphia, 1996; Chapter 30, pp 981−983. (c) Anderson, C. W.; Young, M. E.; Flint, S. J. *Virology* **1989**, *172*, 506−512.

[38] (a) Shaw, J. G. In *Fundamental Virology*, 3rd ed.; Fields, B. N., Knipe, P. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizman, B., Straus, S. E., Eds.; Lippincott-Raven: Philadelphia, 1996; Chapter 12, pp 378−382. (b) Solis, I.; Garcia-Arenal, F. *Virology* **1990**, *177*, 553−558.

[39] (a) Fiers, W.; Contreras, R.; Duerinck, F.; Haegeman, G.; Iserentant, D.; Merregaert, J.; Jou, W. M.; Molemans, F.; Raeymaekers, A.; Van Den Berghe, A.; Volckaert, G.; Ysebaert, M. *Nature* **1976**, *260*, 500−507. (b) Campbell, A. M. In *Fundamental Virology*, 3rd ed.; Fields, B. N., Knipe, P. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizman, B., Straus, S. E., Eds.; Lippincott-Raven: Philadelphia, 1996; Chapter 15, p 463.

[40] Eriksson, J.; Chait, B. T.; Fenyo, D. *Anal. Chem.* **2000**, *72*, 999−1005.

Table 3. Virus Identification Results by Querying a Specially Constructed Database

(a) Two Minutes

| | | | observed peptide mass (Da) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| virus | score[a] | α value | 1497.0 | 1584.7 | 1741.7 | 1766.9 | 1805.4 | 1860.1 | 1975.2 | 2233.6 | 2242.9 | 2264.3 | 2354.1 |
| **SIN** | 20/22 | $2.3 \times 10^{-6}$ | x | x | x | x | x | x | x | x | x | x | x |
| **LV** | 17/22 | $6.6 \times 10^{-3}$ | | x | x | | x | x | x | | x | x | x |
| **RRV** | 15/22 | $3.5 \times 10^{-3}$ | x | | x | x | x | x | x | | x | x | |
| **AD5** | 15/22 | $3.5 \times 10^{-3}$ | x | x | x | x | x | x | x | | | x | x |
| **TMV** | 13/22 | $1.2 \times 10^{-2}$ | x | x | x | x | | x | x | | | | x |
| **MS2** | 4/22 | $5.8 \times 10^{-1}$ | | | | | | | x | | | x | |

| | observed peptide mass (Da) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| virus | 2358.7 | 2398.5 | 2503.6 | 2793.5 | 2983.2 | 2993.5 | 3043.5 | 3196.2 | 3200.9 | 3636.7 | 3717.0 |
| **SIN** | x | x | | | x | x | x | x | x | x | x |
| **LV** | | x | x | x | x | x | x | x | | x | x |
| **RRV** | x | x | x | | x | x | x | | x | | |
| **AD5** | x | x | x | x | | | x | x | | | |
| **TMV** | | x | | x | x | x | x | x | | | |
| **MS2** | | | | x | | | x | | | | |

(b) Five Minutes

| | | | observed peptide mass (Da) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| virus | score[a] | α value | 1245.7 | 1496.8 | 1585.0 | 1744.1 | 1766.5 | 1808.2 | 1859.7 | 1978.7 | 2133.9 | 2138.6 | 2235.8 |
| **SIN** | 21/21 | $4.1 \times 10^{-7}$ | x | x | x | x | x | x | x | x | x | x | x |
| **LV** | 18/21 | $2.9 \times 10^{-5}$ | x | | x | x | | x | x | x | x | | x |
| **RRV** | 17/21 | $5.0 \times 10^{-3}$ | x | | | x | x | x | x | x | x | x | x |
| **AD5** | 15/21 | $1.9 \times 10^{-2}$ | x | x | x | x | x | x | x | x | x | | |
| **TMV** | 9/21 | $2.4 \times 10^{-1}$ | x | x | x | x | x | | x | | | | x |
| **MS2** | 6/21 | $8.5 \times 10^{-2}$ | | | x | | | | | x | x | | |

| | observed peptide mass (Da) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| virus | 2242.9 | 2264.8 | 2352.4 | 2507.6 | 2985.4 | 3041.1 | 3217.8 | 3224.7 | 3638.6 | 3714.4 |
| **SIN** | x | x | x | x | x | x | x | x | x | x |
| **LV** | x | x | x | x | x | x | x | x | x | x |
| **RRV** | x | x | | | x | x | x | x | x | x |
| **AD5** | | x | x | x | | x | x | x | | |
| **TMV** | | | x | x | | | | | | |
| **MS2** | | x | | | x | | | | | x |

[a] A hit occurs when the mass difference between the observed peak and the *in silico* generated peak is less than 2 Da.

The probability of false matches, i.e., the α value, can be calculated by the following equations:

$$\alpha = \sum_{k=k_{\text{match}}}^{K} P_K(k)$$

where

$$P_K(k) = \frac{1}{K\sigma\sqrt{2\pi}} e^{-((k/K - S_0)^2)/(2\sigma^2)}$$

$$\sigma = \sqrt{(S_0 e^{-k/K})/K}$$

$$S_0 = 1 - e^{-n/n^*}$$

$$n^* = (m_{\max} - m_{\min})/\Delta m$$

In the above equations, $\Delta m$ is the mass tolerance and $n^*$ is a parameter—critical distribution density—depending on the mass tolerance. Its physical meaning reflects the maximum number of tryptic peptides from an individual virus that can be contained in the database. For $\Delta m = 2$ Da and for the chosen mass interval, $n^*$ is equal to 1500 peptides. As already pointed out,[16] database pruning, e.g., including tryptic peptides only for proteins expressed in high copy number, can accommodate even the largest viral genomes. Without database pruning, the maximum number of tryptic peptides for microorganisms with larger genomes will exceed the critical distribution density used here and increased mass accuracy will be required to increase the critical distribution density value. For instance, in an unrelated study of all tryptic peptides from an adenovirus by MALDI Fourier transform ion cyclotron resonance mass spectrometry,[42] the achieved $\Delta m$ is better than 0.01 Da (which corresponds to a critical density $n^*$ of more than 300 000 peptides for the range from 1000 to 4000 Da). The maximum number of proteolytic peptides for microorganisms can also be reduced by use of residue specific proteolytic reactions that produce fewer (larger) peptide products.

(41) Fenyo, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998−1005.

(42) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 2836−2842.

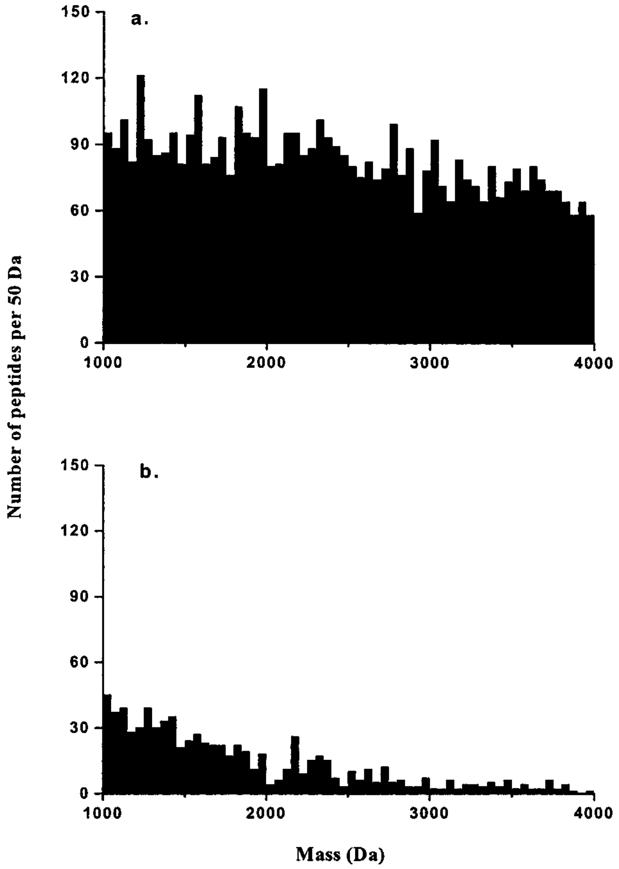

Figure 3. Combined molecular mass distribution of all tryptic peptides from all proteins for the six different viruses: (a) four missed tryptic cleavage sites allowed (4980 peptides total) and (b) no missed cleavage sites allowed (784 peptides total).

The calculated α values (significance factors) for the experimentally observed spectra are also listed in Table 3. The enhanced discriminatory power of the significance testing algorithm is illustrated by the 100-fold difference in the α values for identification of SIN from that of the next closest candidates.

## CONCLUSION AND PROSPECTS

The strategy introduced here allows identification of viral samples in less than 5 min, in a MALDI-TOF mass spectrometer with limited resolution including field-portable units. The reliability of this identification is enhanced by the use of a significance testing algorithm, which in turn requires that only a limited number of proteolytic peptides is experimentally observed. When trypsin is used, this means that proteolysis must be limited. This is not difficult to achieve when the analyte is an intact virus and the incubation time is of the order of minutes. Observation of incompletely cleaved peptides improves the analysis since their molecular masses are spread more evenly across the range of detection. In the future, construction of an on-line tryptic peptide database for all viruses with known genome sequences is envisioned. That database and query algorithms will incorporate additional information, such as protein copy number, posttranslational modifications, virus topology, etc. On the experimental side, proteolytic digestion of viruses directly on the MALDI slide will further improve the sensitivity and simplicity of the method. Implementation of orthogonal approaches—e.g., MS/MS of individual tryptic peptides for generation of sequence tag information—will improve the identification capability, particularly in cases where more than one viral organism is present. Experiments to extend the method to other microorganisms and other proteolytic enzymes are planned as well.

## ACKNOWLEDGMENT

We thank Amy Freas (University of Maryland) for growing the microorganisms used in this study. Financial support for this research was provided by DARPA. Reported at the 49th ASMS Conference on Mass Spectrometry and Allied Topics, Chicago IL, May 28-31, 2001

Received for review January 9, 2002. Accepted March 15, 2002.

AC0200217