Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 1999

Impact Factor: 5.64 · DOI: 10.1021/ac9810516 · Source: PubMed

3 AUTHORS:

Karl R Clauser
Broad Institute of MIT and Harvard
67 PUBLICATIONS **5,057** CITATIONS
SEE PROFILE

Peter R Baker
University of California, San Francisco
26 PUBLICATIONS **1,736** CITATIONS
SEE PROFILE

Alma L Burlingame
University of California, San Francisco
622 PUBLICATIONS **26,615** CITATIONS
SEE PROFILE

Anal. Chem. **1999,** *71,* 2871−2882

Role of Accurate Mass Measurement (±10 ppm) in Protein Identification Strategies Employing MS or MS/MS and Database Searching

Karl R. Clauser,[†,‡] Peter Baker,[†] and Alma L. Burlingame*[,†,§]

Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-0446, and Ludwig Institute for Cancer Research and Department of Biochemistry, University College London, London, U.K.

We describe the impact of advances in mass measurement accuracy, ±10 ppm (internally calibrated), on protein identification experiments. This capability was brought about by delayed extraction techniques used in conjunction with matrix-assisted laser desorption ionization (MALDI) on a reflectron time-of-flight (TOF) mass spectrometer. This work explores the advantage of using accurate mass measurement (and thus constraint on the possible elemental composition of components in a protein digest) in strategies for searching protein, gene, and EST databases that employ (a) mass values alone, (b) fragment-ion tagging derived from MS/MS spectra, and (c) de novo interpretation of MS/MS spectra. Significant improvement in the discriminating power of database searches has been found using only molecular weight values (i.e., measured mass) of >10 peptide masses. When MALDI-TOF instruments are able to achieve the ±0.5−5 ppm mass accuracy necessary to distinguish peptide elemental compositions, it is possible to match homologous proteins having >70% sequence identity to the protein being analyzed. The combination of a ±10 ppm measured parent mass of a single tryptic peptide and the near-complete amino acid (AA) composition information from immonium ions generated by MS/MS is capable of tagging a peptide in a database because only a few sequence permutations >11 AA's in length for an AA composition can ever be found in a proteome. De novo interpretation of peptide MS/MS spectra may be accomplished by altering our MS-Tag program to replace an entire database with calculation of only the sequence permutations possible from the accurate parent mass and immonium ion limited AA compositions. A hybrid strategy is employed using de novo MS/MS interpretation followed by text-based sequence similarity searching of a database.

Rapid progress in genome sequencing of several model organisms is producing a vast sequence infrastructure to facilitate biological and medical research,[1] as well as diagnosis of disease. Proteomic approaches take advantage of this infrastructure to elucidate the biological function/dysfunction of whole suites of proteins comprising the molecular machinery of cellular processes.[2] Success in the push toward automated high-throughput strategies for the characterization of the proteome by mass spectrometric technologies will require significant improvements in the integration and automation of protein isolation and separation techniques, sample delivery to the mass spectrometer, spectral acquisition, and spectral interpretation/database searching. The overall quality of data produced in such studies will be largely determined by the ability of isolation techniques to indicate differential protein expression and the certainty of protein identification resulting from the combination of mass spectrometry and spectral interpretation/database search software. The ability to identify a protein with a high probability of being correct hinges primarily on two disparate aspects of mass spectral quality: mass measurement accuracy in MS experiments and the degree of peptide sequence completeness derived from MS/MS experiments.

In 1954, Beynon and co-workers[3] demonstrated that the measurement of molecular weight accurately determines the unique elemental composition of an organic substance based on permutation and matching with the known noninteger atomic mass values of the common elements. In those experiments, a Nier-type double-focusing magnetic sector instrument was required to obtain the necessary accuracy in mass measurement. This technique is still in routine use today.[4] When MALDI was discovered and implemented on linear TOF mass spectrometers,[5] the accuracy of mass measurement was extremely poor by comparison, i.e., ±2 Da, for peptides using external calibration. This poor accuracy and poor mass resolution [<500 ($M/\Delta M$, fwhm)] were largely caused by formation of ions with both broad initial kinetic energy distributions[6] and mass-independent initial velocities.[7,8] Refocusing of ions using an electrostatic mirror or reflector boosted the achievable resolution to >2000 ($M/\Delta M$, fwhm). However, because the mass distributions of peptides

* To whom correspondence should be addressed at the University of California. E-mail: alb@itsa.ucsf.edu.

† University of California.

‡ Current address: Department of Protein Technologies, Millennium Pharmaceuticals, Inc., 40 Erie St., Cambridge, MA 02139. E-mail: clauser@mpi.com.

§ University College London.

(1) Lander, E. S. *Science* **1996**, *274*, 536−539.
(2) Alberts, B. A. *Cell* **1998**, *92*, 291−294.
(3) Beynon, J. H. *Nature* **1954**, *174*, 735.
(4) Biemann, K. *Methods Enzymol.* **1990**, *193*, 295−305.
(5) Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes* **1987**, *78*, 53−68.
(6) Zhou, J.; Ens, W.; Standing, K. G.; Verentchikov, A. *Rapid Commun. Mass Spectrom.* **1992**, *6*, 671−678.
(7) Beavis, R. C.; Chait, B. T. *Chem. Phys. Lett.* **1991**, *181*, 479−484.
(8) Pan, Y.; Cotter, R. J. *Org. Mass Spectrom.* **1992**, *27*, 3−8.

10.1021/ac9810516 CCC: $18.00
Published on Web 05/27/1999

<2500 Da are only ~0.3 Da wide for each nominal mass unit,[9] in practice reflectors typically only enabled the measurement of the mass of peptides to the correct nominal value. Revival and implementation of the 40-year-old Wiley−McLaren theories for time-lag focusing[10] or delayed extraction (DE) alleviated the ion energy spread problem plaguing MALDI.[11−14] DE accompanied by the modern high-voltage electronic switches and fast digitization electronics (500 MHz−2 GHz) not available in Wiley and McLaren's time now enable DE-MALDI-TOF instruments with reflectors and suitable flight tube lengths to achieve resolution >10 000 ($M/\Delta M$, fwhm) with accompanying mass accuracy (~10 ppm) that approaches the practical limits dictated by elemental composition.[13,15] Furthermore, Fourier transform ion cyclotron resonance instruments which can achieve significantly higher resolution >100 000 ($M/\Delta M$, fwhm) have recently attained even better mass accuracy (~1 ppm)[16] and been applied to study the completeness, diversity, and degeneracy of peptide combinatorial libraries.[17]

In 1989, the potential for identifying proteins by searching sequence databases using the peptide masses measured following enzymatic digestion of an isolated protein was described by Henzel et al.[18] The approach, now commonly referred to as "peptide mass fingerprinting", was implemented by five independent groups who developed database search software and used linear MALDI-TOF instruments.[19−23] However, in practice, studies to identify proteins in this manner rarely relied on the peptide masses alone. Most investigators, including developers of peptide mass fingerprinting software, typically supplemented the peptide mass data with partial sequence data obtained primarily from MS/MS experiments. Using linear or reflector MALDI-TOF instruments without DE, peptide mass fingerprints alone rarely yielded false-positive identifications. Instead, they simply did not routinely yield unambiguous answers with high confidence levels. To improve discriminating power, database searches with masses alone were typically restricted by additional parameters such as protein MW and species of origin, which are both typically known.

Led by some of the early developers of peptide mass fingerprinting software, several groups moved on to the obvious next step of developing database search software that used the partial sequence information present in peptide MS/MS spectra.[24−28] The discriminating power of sequence for searching databases is sufficiently high that the partial sequence in an MS/MS spectrum of a single peptide is adequate for high-confidence protein identification, if the identical sequence or one very similar corresponding to the peptide studied is present in the database. Although sequencing of peptides by MS/MS was first described in 1986,[29] widespread use of MS/MS for that task has been hindered by the cost of instrumentation, the skill required for both instrument operation and spectral interpretation, and the degree of incompleteness or ambiguity in the interpreted sequences. The determination of complete sequences by MS/MS has primarily been possible only with tandem mass spectrometers employing high-energy collision-induced dissociation (CID): four-sector[30,31] and hybrid sector/TOF[32−34] instruments. This situation is destined to change soon with the development of tandem TOF technology.[35] Low-energy CID instruments (triple quadrupole,[29] hybrid sector−quadrupole,[36] and ion traps[37]) typically produce MS/MS spectra yielding partial sequences of varying degrees of completeness. Reflector TOF instruments which exploit metastable fragmentation,[38] in our experience, produce postsource decay (PSD) also yielding partial sequences.[26,32,39−41] While rich in immonium ions, PSD spectra are typically more difficult to interpret de novo than low-energy CID spectra because of the preponderance of internal acylium ions and sequence ions arising from more discontinuous cleavage of the peptide backbone, often at nonconsecutive amino acid positions. Prior to the advances in high-throughput genome sequencing techniques, the skill-level barrier for MS/MS spectral interpretation was reduced by means of software algorithms that derived de novo sequences from the spectra by brute-force

(9) Mann, M. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry & Allied Topics, Atlanta, GA, May 21−26, 1995*; p 639.
(10) Wiley: W. C.; McLaren, I. H. *Rev. Sci. Instrum.* **1955**, *26*, 1150−1157.
(11) Colby, S. M.; King, T. B.; Reilly, J. P. *Rapid. Commun. Mass Spectrom.* **1994**, *8*, 865−868.
(12) Brown, R. S.; Lennon, J. J. *Anal. Chem.* **1995**, *67*, 1998−2003.
(13) Vestal, M. L.; Juhasz, P.; Martin, S. A. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1044−1050.
(14) Juhasz, P.; Vestal, M. L. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 892−911.
(15) Edmondson, R. D.; Russell, D. H. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 995−1001.
(16) Guan, S.; Marshall, A. G.; Scheppele, S. E. *Anal. Chem.* **1996**, *68*, 46−71.
(17) Nawrocki, J. P.; Wigger, M.; Watson, C. H.; Hayes, T. W.; Senko, M. W.; Benner, S. A.; Eyler, J. R. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1860−1864.
(18) Henzel, W. J.; Stults, J. T.; Watanabe, C. Paper presented at the ASMS Conference, Seattle, WA, July 29−Aug 2, 1989.
(19) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.
(20) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−345.
(21) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.
(22) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.
(23) Yates, J. R. D.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.
(24) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(25) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(26) Clauser, K. R.; Baker, P. R.; Burlingame, A. L., *Proceedings of the 44th ASMS Conference on Mass Spectrometry & Allied Topics, Portland, OR, May 12−16, 1996*; p 365.
(27) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(28) Fenyo, D.; Chait, B. Database software, http://prowl.rockefeller.edu.
(29) Hunt, D. F.; Yates, J. R. D.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−6237.
(30) Scoble, H. A.; Biller, J. E.; Biemann, K. B. *Fresnius Z. Anal. Chem.* **1987**, *327*, 239−245.
(31) Medzihradszky, K. F.; Burlingame, A. L. *Methods: A Companion to Methods in Enzymology*; Academic Press: London, 1994; Vol. 6, pp 284−303.
(32) Medzihradszky, K. F.; Adams, G. W.; Bateman, R. H.; Green, M. R.; Burlingame, A. L. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 1−10.
(33) Medzihradszky, K. F.; Maltby, D. A.; Qiu, Y.; Yu, Z.; Hall, S. C.; Chen, Y.; Burlingame, A. L. *Int. J. Mass Spectrom. Ion Processes* **1997**, *160*, 357−369.
(34) Qiu, Y.; Burlingame, A. L.; Benet, L. Z. *Drug Metab. Dispos.* **1998**, *26*, 246−256.
(35) Vestal, M.; Juhasz, P.; Hines, W.; Martin, S. In *Mass Spectrometry in Biology and Medicine*; Burlingame, A. L., Carr, S. A., Baldwin, M. A., Eds.; Humana Press: Totowa, NJ; in press.
(36) Bean, M. F.; Carr, S. A.; Thorne, G. C.; Reilly, M. H.; Gaskell, S. J. *Anal. Chem.* **1991**, *63*, 1473−1481.
(37) Jonscher, K. R.; Yates, J. R. D. *Anal. Biochem.* **1997**, *244*, 1−15.
(38) Kaufmann, R.; Kirsch, D.; Spengler, B. *Int. J. Mass Spectrom. Ion Processes* **1994**, *131*, 355−385.
(39) Stimson, E.; Truong, O.; Richter, W. J.; Waterfield, M. D.; Burlingame, A. L. *Int. J. Mass Spectrom. Ion Processes* **1997**, *169/170*, 231−240.
(40) Qiu, Y.; Benet, L. Z.; Burlingame, A. L. *J. Biol. Chem.* **1998**, *273*, 17940−17953.
(41) Jimenez, C. R.; Huang, L.; Qiu, Y.; Burlingame, A. L. *Curr. Protocols Protein Sci.*, in press.

combinatorics,[42−45] mimicking of manual interpretation methods,[30,46−48] or applying mathematical graph theory.[49−51] These approaches were overtaken in popularity by the simpler and highly effective algorithmic strategies that leverage the burgeoning genomic databases using either preinterpreted[24] or uninterpreted MS/MS spectra.[25,26,28]

All the database search strategies, including the algorithm developed in our own laboratory,[26] are akin to looking up the answer to an odd-numbered problem in the back of the book. While genomic sequencing will eventually reduce the protein world to one where most of the problems are essentially odd-numbered, investigators who study organisms other than those targeted by genome sequencing strategies will continue to view the world as largely composed of even-numbered problems where the answer-key has not yet been written. Hence it seems reasonable that, for studying unknown proteins, either of two approaches will be followed, perhaps both: (1) use of database-oriented tandem mass spectral interpretation strategies which can be successful in instances of strong homology (1 AA mismatch/peptide);[24,52] (2) matching through weak homology (1−3 AA mismatch/peptide) with sequences already present in the database by linking de novo MS/MS spectral interpretation with programs such as FASTA[53] or BLAST[54] that employ sequence similarity matching.[27] Should homology correlation fail, one would then use the sequences derived de novo from MS/MS spectra to guide PCR-based cloning.[31,55,56]

In the present work, we seek not only to extend initial work on protein identification with masses alone measured by DE-MALDI-TOF from ±30 ppm mass accuracy[57] to ±10 ppm but also to examine the potential for homology-tolerant peptide mass fingerprinting at mass accuracies approaching elemental composition levels using our MS-Fit program. Further, we describe protein identification via database searches employing immonium ion tagging for a single peptide using AA compositions from the low-mass region of high-energy CID and PSD spectra in conjunction with an accurate parent mass, but with little sequence knowledge. We also describe mass accuracy driven modification of our database search program, MS-Tag, for de novo MS/MS spectral interpretation, followed by database searching for homologous sequences using our amino acid substitution-tolerant sequence similarity matching program, MS-Edman.

EXPERIMENTAL SECTION

Protein Purification and Isolation. Apolipoprotein A-1 and protein disulfide isomerase were purified from placental explants and isolated by 2D-PAGE as described elsewhere.[58,59]

In-Gel Trypsin Digestion. Proteins in 2D-PAGE spots were subjected to a previously described[60] seven-step in-gel digestion procedure that included (1) gel maceration, (2) destaining, (3) drying, (4) rehydration with a trypsin buffer solution, (5) incubation for 12−16 h at 37 °C, (6) peptide extraction, and (7) partial salt removal. Reduction and alkylation were not performed; we commonly find acrylamide-modified cysteine residues.[61,62] To minimize adsorptive sample loss, all manipulations were performed in siliconized microfuge tubes.

Mass Spectrometry. Portions (typically 1/15th) of unseparated in-gel tryptic digests were cocrystallized in a matrix of α-cyano-4-hydroxycinnamic acid and analyzed using a PerSeptive Biosystems Voyager Elite XL MALDI-TOF mass spectrometer equipped with delayed extraction and operated in the reflector mode. An accelerating voltage of 25 kV and extraction delay of 100 ns were used. Flight time measurements were sampled with a 2 GHz oscilloscope to ensure that individual peaks in the mass spectra in the MS mode were defined by more than 10 data points. Spectra were internally calibrated using trypsin autoproteolysis peptides. Timed ion selection was employed to selectively transmit an individual peptide and its metastable fragment ions to the reflectron for postsource decay (PSD) sequencing. In the PSD mode, data sampling was done at 500 or 250 MHz. PSD was performed by making 9−12 steps of the reflectron voltage, and for each individual step the voltage was reduced to ∼70% of the previous step. Collision gas (air) was employed (source pressure 1×10^{-6} Torr) for collision-induced dissociation (CID) in steps <200 Da. Segments from each individual step were then stitched together to produce the complete spectrum. The spectral segments acquired for each step were typically obtained from separate positions on the sample target, and data from 50−200 laser shots were collected in each step.

Databases and Search Software. The protein translation of Genbank (Genpept release 98) and a larger nonredundant database compiled from a combination of several publicly available protein databases (NCBInr 3/30/97) were obtained as ASCII text files in the FASTA format from the National Center for Biotech-

(42) Sakurai, T.; Matsuo, T.; Matsuda, H.; Katakuse, I. *Biomed. Mass Spectrom.* **1984**, *11*, 396−399.
(43) Ishikawa, K.; Niwa, Y. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 373−380.
(44) Hamm, C. W.; Wilson, W. E.; Harvan, D. J. *CABIOS* **1986**, *2*, 115−118.
(45) Zidarov, D.; Thibault, P.; Evans, M. J.; Bertrand, M. J. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 13−16.
(46) Siegel, M. M.; Bauman, N. *Biomed. Environ. Mass Spectrom.* **1988**, *15*, 333−343.
(47) Johnson, R. J.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945−957.
(48) Yates, J. R.; Griffin, P. R.; Hood, L. E.; Zhou, J. X. In *Techniques in Protein Chemistry II*; Villafranca, J. J., Ed.; Academic Press: New York, 1991; pp 477−485.
(49) Bartels, C. *Biomed. Environ. Mass. Spectrom.* **1990**, *19*, 363−368.
(50) Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 326−336.
(51) Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V. *CABIOS* **1995**, *11*, 427−434.
(52) Clauser, K. R.; Baker, P. R.; Burlingame, A. L. In preparation.
(53) Pearson, W. R.; Lipman, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(54) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−410.
(55) Brown, J. D.; Hann, B. C.; Medzihradszky, K. F.; Niwa, M.; Burlingame, A. L.; Walter, P. *EMBO J.* **1994**, *13*, 4390−400.
(56) Wen, D. X.; Livingston, B. D.; Medzihradszky, K. F.; Kelm, S.; Burlingame, A. L.; Paulson, J. C. *J. Biol. Chem.* **1992**, *267*, 21011−21019.
(57) Jensen, O. N.; Podtelejnikov, A.; Mann, M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1371−1378.
(58) Genbacev, O.; Joslin, R.; Damsky, C. H.; Polliotti, B. M.; Fisher, S. J. *J. Clin. Invest.* **1996**, *97*, 540−550.
(59) Hoang, V.; Clauser, K. R.; Foulk, R. A.; Genbacev, O.; Zhou, O.; Fisher, S. J.; Burlingame, A. L. In preparation.
(60) Matsui, N. M.; Smith, D. M.; Clauser, K. R.; Fichmann, J.; Andrews, L. E.; Sullivan, C. M.; Burlingame, A. L.; Epstein, L. B. *Electrophoresis* **1997**, *18*, 409−417.
(61) Hall, S. C.; Smith, D. M.; Masiarz, F. R.; Soo, V. W.; Tran, H. M.; Epstein, L. B.; Burlingame, A. L. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1927−1931.
(62) Clauser, K. R.; Hall, S. C.; Smith, D. M.; Webb, J. W.; Andrews, L. E.; Tran, H. M.; Epstein, L. B.; Burlingame, A. L. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 5072−5076.

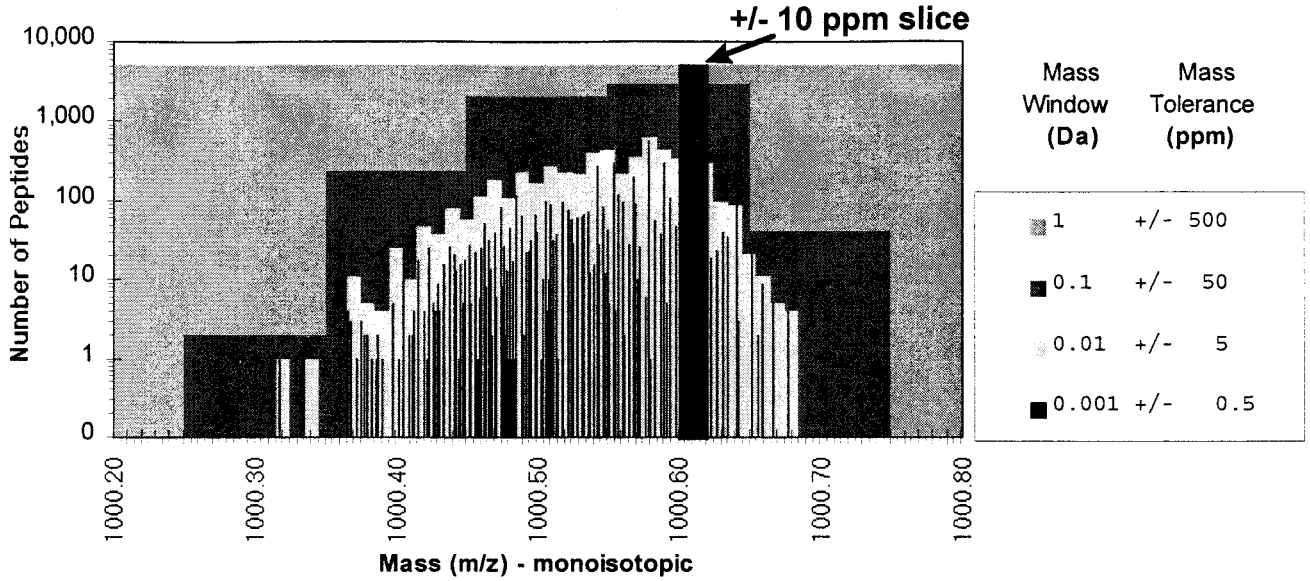

Figure 1. Number of tryptic peptides of nominal mass 1000 present in the Genpept database release 98 (12/96), which contains ~210 000 entries. To facilitate assessment of the discriminating power of mass accuracy, the number of peptides was counted in mass windows of four different widths. One missed cleavage was allowed.

nology Information (NCBI), Washington, DC (ftp://ncbi.nlm.nih.gov/genbank/genpept.fsa.Z and ftp://ncbi.nlm.nih.gov/blast/db/nr.Z), and were used for the studies described here. Both MS-Fit and MS-Tag searches were performed with the following general parameters: protein molecular mass range of 1000−100 000 Da, all species allowed, one missed cleavage allowed for trypsin digests, cysteines modified by acrylamide, and parent ion mass tolerance of ±10 ppm (unless noted otherwise). Additional MS-Tag search parameters included the following: fragment ion mass tolerance of 1000 ppm and allowed fragment ion types a, b, y, a − NH_3, b − NH_3, y − NH_3, b − H_2O, b + H_2O, and internal. The de novo interpretation mode is invoked in MS-Tag by designating Unknome as the database. Our software can be applied to databases in FASTA format. Our worldwide web server (http://prospector.ucsf.edu) also allows searching of the OWL and Swiss-Prot protein databases as well as 6-frame translation of the dbEST DNA database.

RESULTS AND DISCUSSION

Mass Accuracy and the Peptide Limit of Elemental Composition. Improvements in the accuracy of mass measurement are of significant practical value down to the level at which the elemental compositions of possible sample components can be distinguished. For tryptic peptides of mass <2000 Da, this elemental composition limit is reached at mass accuracies of 2−0.5 ppm. Figure 1 illustrates the number of tryptic peptides of nominal mass 1000 Da in a public sequence database and shows in subsets of several different accuracies the number of peptides among which a parent mass measurement alone is unable to discriminate. If the enzyme specificity is ignored, so that all peptides in the database are considered, then the shapes of the distributions in Figure 1 are little changed (data not shown). The number of peptides in each subset simply increases by approximately 100-fold. The absence of any peptides at certain discrete mass values at ±0.5 ppm accuracy indicates that the elemental composition limit is reached. To have a mass at these apparently forbidden mass values would require an elemental composition that is not consistent with a peptide composed of the 20 common amino acids.

Peptide Mass Fingerprinting and Matching Homologous Proteins. The DE-MALDI-TOF mass spectrum shown in Figure 2 was obtained from the peptides recovered from the unseparated in-gel tryptic digestion of a protein isolated from human placental cells. Table 1 illustrates that by masses alone the protein is readily identified as the bovine (cow) form of apolipoprotein A-1 in a database search with MS-Fit. Table 2 shows the mass accuracy associated with each of the 18 peptide masses belonging to an apolipoprotein A-1 peptide measured in Figure 2. We believe the bovine protein to have originated in the human placental cell preparation from fetal calf serum used early in cell culture, although its differential uptake is probably related to the preeclampsia pregnancy disorder under study.[59]

In establishing this identity, ±10 ppm mass accuracy has played an important role. In Table 1, note that few proteins match to more than 4/23 peptide masses. Hence, the potential of a false-positive result is quite low and the discriminating power of the data is very high. However, Table 3 illustrates the results of searching the database using the same data but changing only the peptide mass tolerance supplied as a search parameter. As one would expect with lower mass accuracies of +1−2 Da (currently typical of linear, MALDI-TOF instruments without DE), there is much more ambiguity present in the search results. Potential-false-positive matches occur for proteins matching as many as 15/23 peptides. With mass accuracies in the range of ±0.5 to ±0.1 Da (currently typical of linear DE or reflector continuous-extraction MALDI-TOF instruments), the level where false-positive matches appear falls to 7/23−8/23. Once one reaches the mass accuracies ±50 ppm (external calibration) or ±10 ppm (internal calibration), currently possible with reflector, DE MALDI-TOF instruments of suitable flight tube length, the level where false-positive matches appear falls to 6/23 and 4/23,

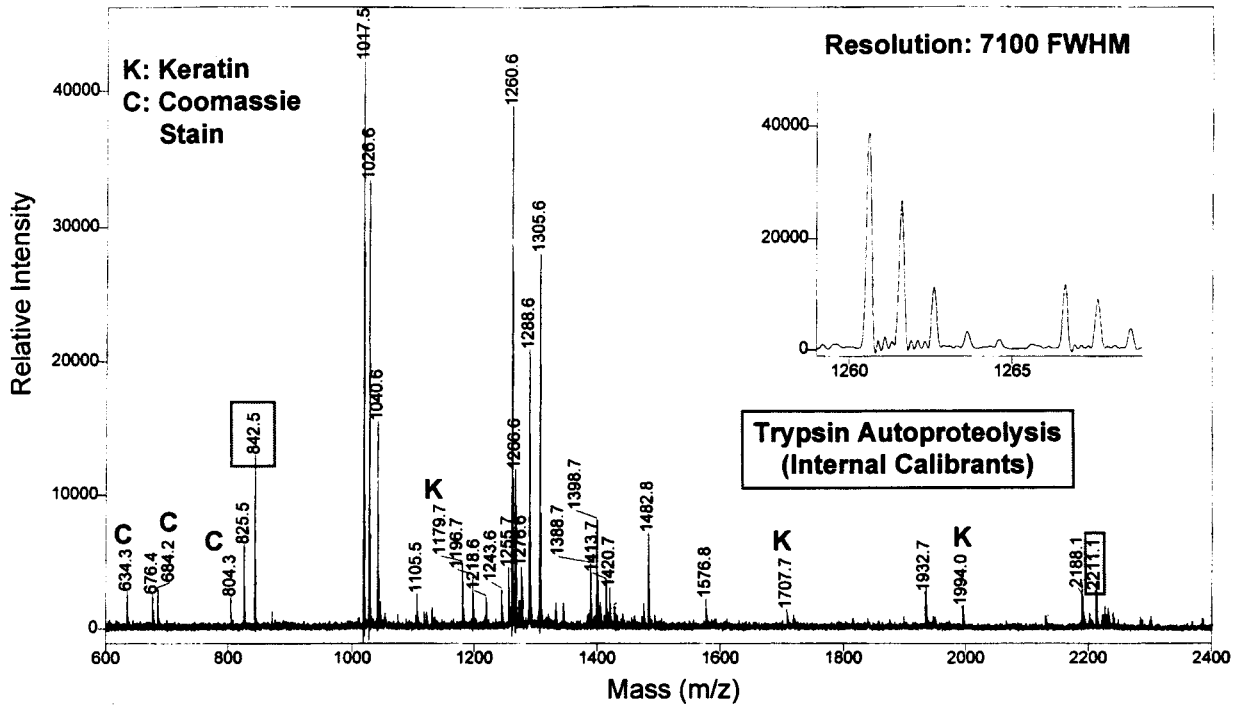

Figure 2. MALDI re-TOF spectrum of an aliquot (1/15th) from the unseparated peptide mixture recovered from in-gel digestion of a 2D-PAGE spot identified in this work as bovine apolipoprotein A-1. Approximately 50−200 fmol of sample was loaded on the MALDI target.

Table 1. MS-Fit Results from a Search[1] Using 23 Masses (not Coomassie, Trypsin, or Keratin) Measured in Figure 2

Rank	#(%)Masses Matched	NCBI gi #	Species	Protein MW (Da)	Protein Name
1	15/23 (65%)	113988	BOS TAURUS	30276.5	apolipoprotein A-I precursor
1	15/23 (65%)	245563	BOS TAURUS	28432.2	apolipoprotein A-I
2	13/23 (56%)	162678	BOS TAURUS	30271.5	apolipoprotein A-I precursor
3	4/23 (17%)	423201	SUS SCROFA	30330.5	apolipoprotein
3	4/23 (17%)	108523	BOS TAURUS	5188.8	apolipoprotein A-I (fragment)
3	4/23 (17%)	117324 6	ACHOLEPLASMA FLORUM	26934.1	ribosomal protein S3
3	4/23 (17%)	155305 3	STREPTOCOCCUS PNEUMONIAE	72926.6	penicillin-binding protein

[1] +/- 10 ppm mass tolerance - monoisotopic , all species, 1-100 kDa protein MW, trypsin digest, 1 missed cleavage allowed, Cys modified by acrylamide, a minimum of 4 masses matched, and considered modifications: peptide N-terminal Gln to pyroGlu, oxidation of Met, protein N-terminus acetylated.

respectively. The level at which false positives appeared prior to the advent of DE led to additional restrictions, such as the intact protein molecular weight and the species, being used to improve the discriminating power of peptide mass fingerprinting searches.[19−23] The dramatically increased discriminating power of <50 ppm mass accuracy also holds considerable promise for identification of the components in simple mixtures of one to three proteins.[63,64]

Interestingly, a homologous protein, the Sus scrofa (pig) form of apolipoprotein A-1 (88% overall sequence identity) appears in the search results in Table 1, while the human form (77% overall sequence identity) does not. Inspection of the bovine and porcine sequences shows that there are three identically conserved tryptic peptides, while the bovine and human sequences share a single conserved tryptic peptide. Closer inspection reveals that an additional eight porcine peptides and four human peptides each have only a single AA mismatch (which changes the peptide mass) with respect to the corresponding bovine peptides. Following a single AA substitution, an additional two porcine and three human peptides have identical mass and elemental composition but a nonidentical yet homologous sequence in relation to the corresponding bovine peptides. These mass spectrometrically fortuitous occurrences stem from the fact that homologous amino acids are typically related by the addition or subtraction of a methylene unit CH_2 (Gly and Ala, Ser and Thr, Val and Ile/Leu, Asp and Glu, Asn and Gln). Hence it occurred to us that because mass accuracy

(63) Jensen, O. N.; Podtelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997**, *69*, 4741−4750.

(64) Zhang, W.; Chait, B. Identification of components in simple mixtures, http://prowl.rockefeller.edu.

Table 2. Mass Accuracy (Internal Calibration) of 18 Bovine Apolipoprotein A-1 Peptides in Figure 2

Mass Measured	Calculated Mass[1]	Delta Mass (ppm)	start	end	Sequence[2]	Modification
676.3718	676.3630	13.0810	83	88	(K)ETASLR(Q)	
825.4571	825.4583	-1.4341	159	165	(R)AHVETLR(Q)	
1017.5370	1017.5369	0.0658	124	132	(K)VAPLGEEFR(E)	
1026.5964	1026.5948	1.6227	146	154	(K)LSPLAQELR(D)	
1040.6121	1040.6104	1.5768	211	219	(K)AKPVLEDLR(Q)	
1196.7132	1196.7254	-10.2359	238	248	(R)QGLLPVLESLK(V)	
1218.5796	1218.5755	3.3721	206	217	(K)EGGGSLAEYHAK(A)	
1255.6538	1255.6574	-2.8906	36	46	(K)DFATVYVEAIK(D)	
1260.5928	1260.6013	-6.7731	113	121	(K)WHEEVEIYR(Q)	
1266.6281	1266.6370	-7.0870	102	111	(K)VQPYLDEFQK(K)	
1288.6094	1288.6174	-6.1932	166	176	(R)qQLAPYSDDLR(Q)	pyroGlu
1305.6310	1305.6439	-9.8980	166	176	(R)QQLAPYSDDLR(Q)	
1388.6975	1388.6963	0.8894	112	121	(K)KWHEEVEIYR(Q)	
1398.6805	1398.6905	-7.1343	33	45	(R)DYVAQFEASALGK(Q)	
1482.8101	1482.8208	-7.2483	16	28	(R)VKDFATVYVEAIK(D)	
1576.8138	1576.8223	-5.3401	51	64	(K)LLDNWDTLASTLSK(V)	
1932.7067	1932.9343	-117.7835[3]	67	82	(R)EQLGPVTQEFWDNLEK(E)	
2188.0850	2188.1039	-8.6362	65	82	(K)VREQLGPVTQEFWDNLEK(E)	

[1] Monoisotopic masses, unattributed masses; 1105.4550, 1243.5887, 1276.5836, 1413.7126, 1420.6741.
[2] () residues preceeding and following peptide.
[3] The mass accuracy deviation at 1932 Da is a known artifact of instrument geometry. With the low mass gate (matrix suppression) set at 500 Da, ions of mass ~1930 are perturbed as they pass the back side of the reflector detector while entering the reflector by the low mass gate "opening" (detector voltage being initialized).

Table 3. MS-Fit Searches[1] at Various Mass Tolerances Using 23 Masses Measured in Figure 2 (Dashed Lines Show Levels Below Which Only the Correct Proteins Are Matched)

Minimum # Peptides Matched	Number of Proteins Matched: Mass Tolerance supplied to MS-Fit ±2.0 Da	±1.0 Da	±0.5 Da	±0.3 Da	±0.1 Da	±50 ppm	±10 ppm
1	156,793	117,419	77,906	77,374	63,730	47,461	11,703
2	104,022	58,188	24,997	24,708	16,842	9,344	723
3	67,400	26,460	7,455	7,297	4,087	1,766	36
4	42,295	11,623	2,048	1,991	923	323	7
5	25,638	4,846	509	496	190	44	3
6	14,987	1,882	145	135	51	8	3
7	8,192	687	36	33	10	3	3
8	4,378	248	12	9	3	3	3
9	2,208	88	3	3	3	3	3
10	1,062	35	3	3	3	3	3
11	466	9	3	3	3	3	3
12	200	3	3	3	3	3	3
13	72	3	3	3	3	3	3
14	34	3	3	3	3	3	2
15	12	3	3	3	3	3	2
16	3	3	3	3	2	2	0

[1] All parameters unchanged from Table 1, except mass tolerance.

reduces the level of false positives so dramatically in peptide mass fingerprint searches, it might be possible to loosen other parameters and make the searches homology-tolerant.

We reasoned that if a single AA substitution was allowed which could convert a homologous peptide in the database to one with the same mass as a peptide being analyzed, then the relative ranking of homologous proteins would improve. Obviously, if one allows single AA substitutions, then the effective size of the database being searched increases and the level of false positives should also rise. Thus the relative rise in search ranking for

Table 4. Homology-Tolerant MS-Fit Search Results Using 18 Calculated Bovine Apolipoprotein A-1 Peptide Masses with a Single AA Substitution/Peptide Allowed[1] (Dashed Lines Show Levels Below Which Only Homologous Proteins Are Matched

Minimum # Masses matched	Proteins Matched: All species 1-100 kDa MW ±10ppm	±5ppm	±1ppm	Mammals[2] 1-100 kDa MW ±10ppm	±5ppm	±1ppm	All species 1-50 kDa MW ±10ppm	±5ppm	±1ppm	Mammals[2] 1-50 kDa MW ±10ppm	±5ppm	±1ppm
1	8,072	4,616	2,064	1,403	849	460	4,927	2,853	1,300	844	537	312
2	7,776	4,262	1,472	1,352	780	301	4,634	2,519	791	794	473	173
3	7,261	3,632	899	1,226	632	170	4,135	1,944	383	673	336	83
4	6,623	2,919	440	1,103	515	88	3,542	1,355	156	557	236	41
5	5,906	2,228	208	989	397	50	2,919	838	58	458	148	24
6	5,064	1,670	102	886	314	24	2,233	537	30	375	108	17
7	4,217	1,124	56	736	212	16	1,604	289	23	265	67	15
8	3,332	715	37	588	138	15	1,052	153	21	177	43	15
9	2,458	381	23	440	77	15	608	74	18	117	24	15
10	1,774	200	19	325	46	15	321	37	18	68	19	15
11	1,158	110	16	218	32	13	146	23	15	37	15	13
12	669	47	15	127	19	12	75	19	14	22	15	12
13	367	25	10	70	15	9	37	18	10	15	15	9
14	160	15	5	38	12	4	19	14	5	14	12	4
15	68	8	3	20	5	3	12	7	3	9	5	3
16	18	3	3	9	3	3	6	3	3	5	3	3

[1] A single AA could be substituted by one of the standard 20 AA 's, or oxidized Met (m), or the N-terminus of a peptide could be substituted by pyro-glutamic acid. Peptide C-terminal substitutions which eliminated an enzyme cleavage site were not allowed.

[2] Mammals was allowed to consist of the following species: human, cat, chimpanzee, cow, dog, goat, gorilla, macaque, mouse, pig, rabbit, rat, and sheep.

homologous proteins must outpace the increase in ranking of false positives, to gain a discriminating power improvement with a homology-tolerant search strategy that allows single AA substitutions.

Table 4 illustrates the results of single AA substitution-tolerant searches associated with variation of several relevant search parameters, when the calculated (not measured in Figure 2) values for the 18 bovine apolipoprotein A-1 peptides observed in Figure 2 are used. Only those single AA substitutions which differ in mass by <40 Da were allowed. A value of 39 Da is the largest mass change associated with a single AA substitution in which the two AA's involved are strongly homologous (Phe to Trp); larger mass changes are only associated with substitutions involving nonhomologous or weakly homologous AA's. For reasons of computer search time associated with making AA substitutions, a preliminary filter was employed which required a candidate protein to match one of the submitted masses without any AA substitutions. Protein MW was evaluated as an obvious source of false positives; larger proteins have more tryptic peptides, and a single AA substitution in one of those peptides is more likely to yield a peptide that is of identical mass yet of dissimilar sequence compared to those of the peptide being analyzed. Furthermore, because the overall level of protein sequence identity should be high for single AA substitutions to yield a peptide of identical sequence, a collection of related species (mammals) was examined in comparison to all species. Table 5 lists the 21 database entries for homologous apolipoprotein A-1 proteins and indicates how the 18 bovine peptide masses can be correlated to the corresponding homologous peptides in each protein.

Taken together, the results in Tables 4 and 5 can lead to a set of guidelines about conditions under which homology-tolerant searches with masses can provide suitable discriminating power for identification: (1) ≥10 tryptic peptides belonging to the protein being studied should be measured. (2) At least one tryptic peptide should be conserved. (3) There should be sufficient mass accuracy to discriminate among peptide elemental compositions (on the order of ±0.5−5 ppm); ±10 ppm is inadequate. (4) The overall sequence identity should be high (>70%), i.e., typical of that found among species in the same phylogenetic order, mammals for example.

If searches are performed with less stringent mass accuracy or species constraints, then the homologous proteins should have intact molecular masses of ~50 kDa or less. Larger proteins yield a greater number of peptides upon enzymatic digestion, thus increasing the probability of a false-positive match of the measured masses.

In practice, when mass spectral data of sufficiently high quality can be obtained to meet these rather restrictive criteria to yield a homology match with masses alone, simple additional experimental effort would likely yield confirmatory partial sequence support by MS/MS. Hence, a single AA substitution search satisfying less strict criteria could be used as a screen to yield a preliminary candidate pool from which MS/MS spectra would serve as the final discriminating tools for identification. This type of strategy is likely to be most useful when one studies proteins derived from an organism for which little genome sequence exists but for which substantial genome sequence from a related organism is available. In work similar to ours, Cordwell and co-workers also recently proposed that peptide mass fingerprinting can identify homologous

Table 5. 16 Homologous Apolipoprotein A-1 Proteins Matched with a Single AA Substitution/Peptide Allowed[1]

Rank	Number Masses Matched[2] Total	0 AA Sub	1 AA Sub[3]	1 AA FP[4]	NCBI gi #	Species	MW	Sequence Identity
1	18	17	1	-	113988	BOS TAURUS	30276.5	100
1	18	17	1	-	245563	BOS TAURUS	28432.2	93.2
2	16	16	-	-	162678	BOS TAURUS	30271.5	99.2
3	14	3	10	1	423201	SUS SCROFA	30330.5	89.1
4	13	1	6	6	178777	HOMO SAPIENS	30763.9	78.7
4	13	1	6	6	178775	HOMO SAPIENS	28961.7	72.3
4	13	1	6	6	113992	HOMO SAPIENS	30778.0	78.7
4	13	1	6	6	490098	HOMO SAPIENS	28078.7	69.9
4	12	2	10	-	113989	CANIS FAMILIARIS	27467.0	76.3
5	12	1	7	4	399042	MACACA FASCICULARIS	30734.9	77.5
5	12	1	7	4	86614	MACACA FASCICULARIS	30718.9	77.2
5	12	3	8	1	461519	SUS SCROFA	30325.5	88.3
6	11	1	9	1	113996	ORYCTOLAGUS CUNICULUS	30591.6	77.4
7	10	1	9	1	1460	ORYCTOLAGUS CUNICULUS	30518.6	76.7
7	10	2	7	1	346458	SUS SCROFA	30254.4	87.2
8	6	2[5]	3	1	231557	MUS MUSCULUS	30587.6	66.6

[1] Same AA substitutions as allowed in Table 4.
[2] Using calculated masses of 18 bovine apolipoprotein A-1 peptides listed in table 2.
[3] Single AA substitution allows peptides of homologous, perhaps non-identical sequence, but identical elemental composition to match input mass.
[4] FP (false-positive), single AA substitution allows dis-similar peptide to match input mass.
[5] 1 mass matches a sequence identical to bovine apolipoprotein A-1, the other is a false-positive match to a dis-similar sequence.

proteins across species boundaries using a ±6 Da peptide mass tolerance, because a few tryptic peptides are completely conserved.[65] However, they show by way of example that even when a search is limited by protein MW to 72 018 proteins, discriminating power is inadequate for confident identification. The major difference in the present work is the separation of mass tolerance into two components, each of which is retained in the database search: (1) accuracy of the peptide mass measurements and (2) mass shift associated with AA substitution.

Theoretical Considerations of Peptide Sequence Ambiguity for the Post-Genomic/Proteomic Era. Only a tiny fraction of the theoretically possible peptide sequences can ever actually be present in the proteome of a living organism. Figure 3 illustrates the numbers involved. The number of theoretically possible sequences increases exponentially with peptide length (N) and can be simply calculated as 20^N. While the actual number of sequences present in a genome spans a single order of magnitude for typical peptide lengths <25, the actual number of peptides of length N is approximated by (the number of genes in a genome) × (average protein length − N). It is important to recognize that these numbers represent the maximum and assume that a particular sequence is unique in a proteome. However, since certain sequence motifs have protein structure and function roles, some sequences would be repeated in a proteome and the actual number of unique sequences would be lower. From these calculations, it is obvious that the sheer combinatorial enormity guarantees that only a tiny fraction of the possible sequences >8 AA's in length can ever be found in the proteome of a biological

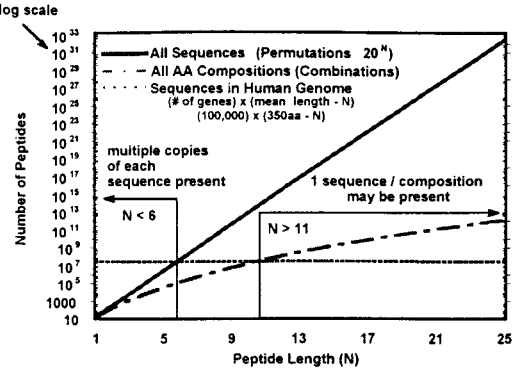

Figure 3. Number of theoretical peptides vs peptide length (N = 1−25). The number of all possible sequences was calculated by 20^N. The number of peptides in the human genome was approximated by (number of genes in genome) × (average protein length − N); number of genes ~ 100 000; average protein length ~ 35 000. The number of all possible AA compositions (no sequence information) was calculated by comp(x, N) = sum{bin($i + x - 2$, i)} + sum{($x - i$) bin($i + N - 2$, i)} for $x > 1$ and $N > 1$; $x = 20$ = number of amino acids; bin(a, b) = $a!/(b!(a - b)!)$; the first sum is from $i = 0$ to $N - 2$; the second sum is from $i = 0$ to $x - 2$ (courtesy of Frank Kragh).

organism. By comparing the values in Figure 3 for sequences in the human genome vs all possible sequences, one finds that for length 7 only 2.7% should be present, while for length 25 only 9.7 × 10^{-24}% should be present. Furthermore, the number of peptide sequences (permutations) of a given length in a proteome is less than the theoretical maximum number of AA compositions (combinations) above length $N = 11$. This means that *not* all the theoretical combinations of AA's for $N > 11$ are represented in a given proteome and that if a peptide's AA composition is

(65) Cordwell, S. J.; Wasinger, V. C.; Cerpa-Poljak, A.; Duncan, M. W.; Humphery-Smith, I. *J. Mass Spectrom.* **1997**, *32*, 370−378.

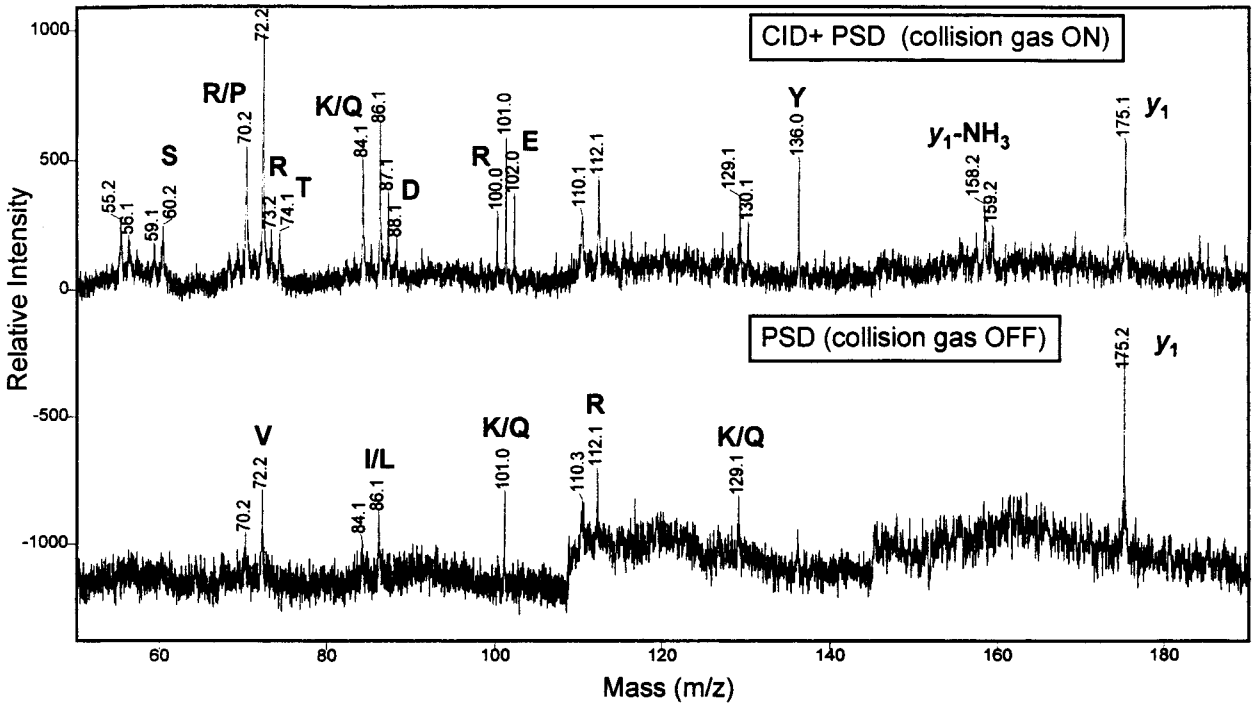

Figure 4. Immonium ion region of the MALDI PSD spectra of the $MH^+ = 1780.8300 \pm 10$ ppm peptide in the unseparated peptide mixture recovered from in-gel digestion of a 2D-PAGE spot identified as VDATEESDLAQQYGVR from protein disulfide isomerase (ER-60). Collision gas (air) was employed. Source pressure was 1×10^{-6} Torr to enhance formation of immonium ions observed under high-energy CID conditions. Approximately 50−200 fmol of sample was loaded on the MALDI target.

determined—using a combination of accurate parent mass and immonium ions for example—there may be only one sequence permutation of that AA composition present in the database. For proteomics studies, this discriminating power provides a tremendous advantage for identification, while studies of combinatorial peptide libraries face a much higher probability that mass and AA composition cannot uniquely identify a peptide.[66]

Immonium Ion Tagging. Information about a peptide's amino acid composition can be determined with a mass spectrometer by accurately measuring the peptide mass and assessing the immonium ion region (50−160 Da) of the peptide's MS/MS spectrum. This region of an MS/MS spectrum contains information on the presence of specific amino acids in the peptide but not on their stoichiometry. The immonium ion region is particularly rich in compositional information when the MS/MS experiment is performed under high-energy conditions that are typical of magnetic sector,[66] hybrid sector/time-of-flight,[32] and time-of-flight instruments.

In the absence of collision gas (air), peptide fragmentation in a reflectron-equipped MALDI-TOF instrument occurs primarily by metastable or postsource decay (PSD) processes that are sequence, matrix, and laser power dependent. Gas collisions impart additional internal energy to the peptide ion. This allows collision-induced dissociation processes in addition to PSD and leads to more extensive fragmentation. Because of the high accelerating voltage (>20 kV) in MALDI-TOF instruments, high-energy (>1 keV) CID processes occur. High-energy CID generally yields more extensive fragmentation and more complete sequence information than PSD or low-energy CID. High-energy CID hallmarks include extensive AA composition information in the immonium ion region, the satellite sequence ion types *d* and *w* (side-chain fragmentation, allows distinguishing I and L), and additional sequence ion types *c*, *x*, *v*, and *z*. Using collision gas in MALDI-TOF instruments typically leads to dramatic improvements in the extent of fragmentation in the immonium ion region. However, if a peptide is already yielding b and y ions by PSD fragmentation, addition of collision gas can result in *d* and *w* ion formation if a basic amino acid such as arginine is present in the ion.[39] The major advantage of adding collision gas for MS/MS by MALDI-TOF is nearly complete AA composition data in the immonium ion region. Figure 4 illustrates the additional immonium and related ions observed when collision gas is added in a PSD experiment to obtain a PSD + CID spectrum.

After measurement of the parent mass of the peptide to 1780.8300 Da ± 10 ppm, PSD + CID yields a partial AA composition of E, V, T, R, S, D, Y, {RP}, {LI}, {NR}, and {KQ}. Values in braces designate that an ion cannot exclusively indicate the presence of a single amino acid but instead limits it to at least one of two possibilities. Using our MS-Comp program (http://prospector.ucsf.edu), this information leads to 875 theoretically possible distinct AA compositions (not sequences) representing 23 distinct elemental compositions. Furthermore, a staggering 1.6×10^{15} possible sequence permutations are possible from these combinations. However, Table 6 shows that searching a database with the immonium ion region spectral information that limits the amino acid composition produces only 2 or 17 matching sequences when the search parameters included or excluded, respectively, the specificity of the enzyme used to generate the peptide from its source protein.

De Novo Peptide Sequence Interpretation and Sequence Similarity Searching. Upon establishing the discriminating

(66) Demirev, P. A.; Zubarev, P. A. *Anal. Chem.* **1997**, *69*, 2893−2900.

Table 6. MS-Tag Search[1] Using 1780.8300 ± 10 ppm Parent Mass and Ions in Figure 4

Sequences passsing parent mass filter					
321 Trypsin			53,372 No Enzyme		
Calculated MH+ (Da)	MH+ Error (ppm)	Sequence	Calculated MH+ (Da)	MH+ Error (ppm)	Sequence
1780.8255	2.5544	(K)GSLAEVQTYDWQNNR(N)	1780.8142	8.8624	(Y)NPGYQDESVLWTESR(D)
1780.8353	-3.0026	(K)VDATEESDLAQQYGVR(G)	1780.8182	6.6035	(I)EDFWSISTYYQVSR(T)
			1780.8255	2.5544	(N)TQLEQVYQGGWNSDR(T)
			1780.8255	2.5544	(W)DYGHSVEQYGALGGTAR(S)
			1780.8255	2.5544	(E)AHEYNDLNTSSVQFR(L)
			1780.8255	2.5544	(K)GSLAEVQTYDWQNNR(N)
			1780.8255	2.5544	(N)TQLEQVYQGGWNSDR(T)
			1780.8288	0.6616	(L)SCYREPGVGEDTQIR(K)
			1780.8353	-3.0026	(K)VDATEESDLAQQYGVR(G)
			1780.8353	-3.0026	(S)YGPEQTTDATDSGLAVR(L)
			1780.8353	-3.0026	(N)TDSYETISGNQVDPVR(L)
			1780.8362	-3.4902	(I)MESCVYHLDVTEKR(V)
			1780.8466	-9.3104	(S)QTNGNLYIANVESSDR(G)
			1780.8466	-9.3104	(G)LKHVTSNASDSESSYR(G)
			1780.8466	-9.3104	(S)TQQDYSPSREVLSDR(I)
			1780.8466	-9.3104	(D)EHCKKIDYGVTCSR(F)

[1] All species, Protein MW 1000 - 200000 Da. Fragment ion types considered: y, Immonium ions for E, V, T, R, S, D, Y, [RP], [LI], [NR], and [KQ]. Brackets indicate limit of ion to at least one of two possibilities.

power of database searches using only the partial AA composition information from an accurate parent mass and the immonium ion region of a peptide MS/MS spectrum, we realized that these same attributes could be powerful attributes in a hybrid strategy employing de novo MS/MS interpretation followed by text-based sequence similarity searching of a database.[53,54] Thus, rather than using the spectral information to search a genome database, it should be possible to generate a database of all possible peptide sequences. Unfortunately, the enormity of permutations (Figure 3) makes this task computationally prohibitive. However, it is a tractable task on conventional personal computers to search only the permutations of the possible AA combinations dictated by an accurate parent mass and immonium ions as shown in Figure 4 and originally described by Zidarov and co-workers.[45] Hence, we have modified MS-Tag to generate on-the-fly a virtual database we call the Unknome, by using immonium ions and the parent mass to construct a restricted set of AA combinations and all permutations of each combination (see Figure 5). The Unknome is searched by MS-Tag, using the fragment ion masses in an MS/MS spectrum in the same way as would be done for a conventional proteome or genome database. In practice, this approach can search portions of the Unknome up to 10 million sequences in a few minutes. This strategy accommodates peptides up to an approximate mass of 1300 Da. Beyond that, search times become prohibitive because of the enormous numbers of sequence permutations that would have to be constructed and examined. Trying to go beyond 1300 Da would probably be best accomplished by replacing the fragment-ion tag algorithm employed in MS-Tag with a more sophisticated mathematical approach which reduces the number of sequence permutations examined. Combinatoric approaches to de novo sequencing based on graph theory have been implemented by several groups.[49−51]

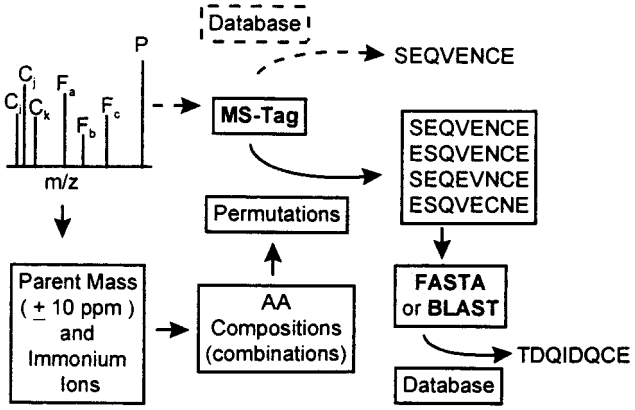

Figure 5. Searching the Unknome. In its original form, our program MS-Tag uses the fragment-ion tag data in an MS/MS spectrum to search a genome database for matches to identical or strongly homologous sequences (dashed lines). It can also be used in a de novo sequencing mode in which the parent mass (P) and composition ions (C_i, C_j, C_k, ...) are used to first calculate all possible AA compositions and then assemble all resulting permutations. This virtual database, which we call the Unknome, is then substituted for a genome database and searched by MS-Tag using the sequence-related fragment ions (F_i, F_j, F_k, ...). As MS/MS spectra will often yield incomplete fragmentation and thus ambiguous sequences, a set of indistinguishable sequences may be interpreted. This Unknome pool may then be leveraged using a genome database and homology-based search algorithms such as FASTA or BLAST to additionally match weakly homologous sequences such as TDQIDQCE.

Figure 6 illustrates a PSD spectrum obtained for the 1040 Da peptide in Figure 2. In an MS-Tag search of the Unknome the parent mass and immonium ions combine to limit the relevant portion of the Unknome, to 21 AA combinations and 2 943 360 permutations representing only three elemental compositions. An

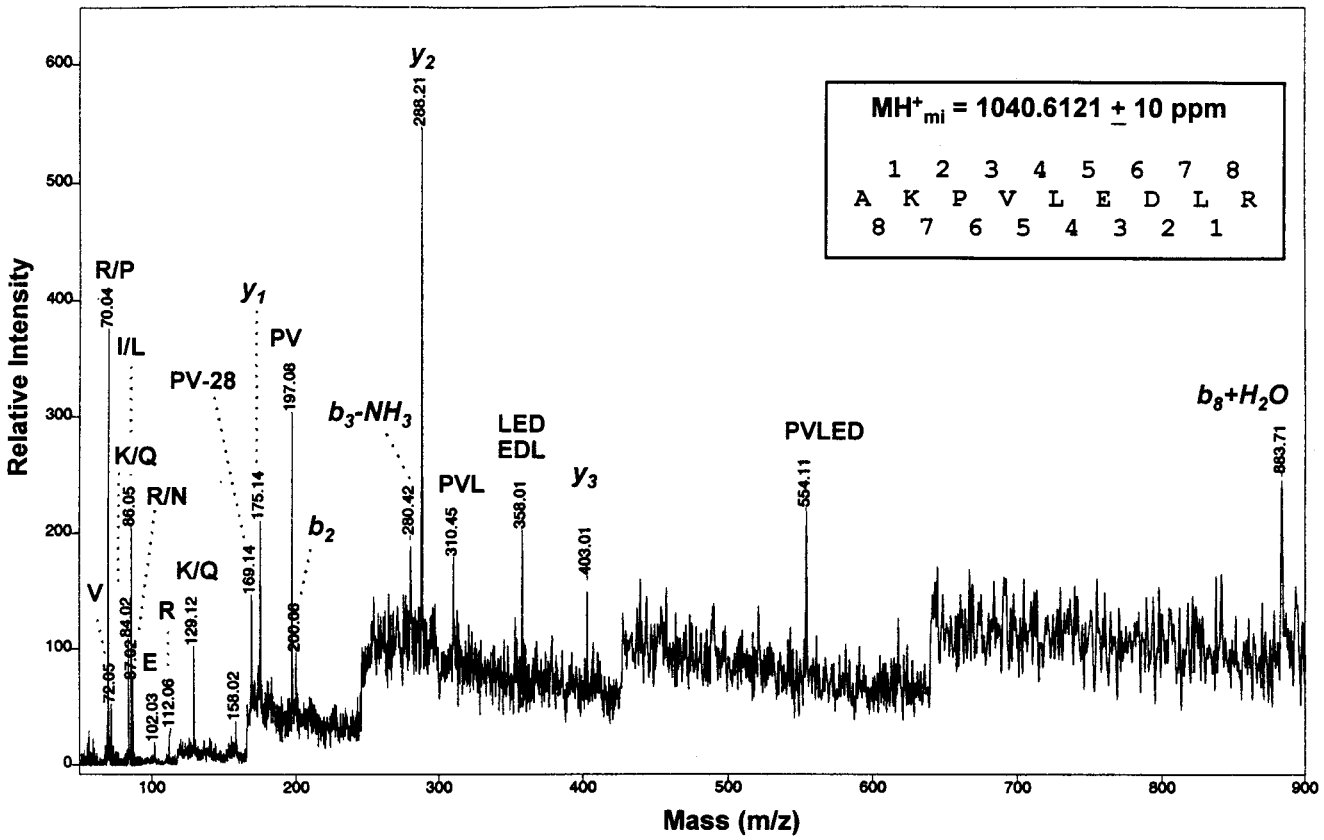

Figure 6. MALDI PSD spectrum of the 1040 Da peptide present in the in-gel tryptic digestion mixture shown in Figure 2. With the timed ion selector (resolving power ~80, $M/\Delta M$) set to 1045 Da, the peptide at mass 1040 Da and its PSD fragments are transmitted to the reflector for fragment-ion mass analysis, while other peptides in the mixture including the 1026 Da peptide and its PSD fragments are excluded. Collision gas (air) was employed during acquistion of the spectral segments below *m/z* 165 to enhance immonium ion formation.

MS-Tag search using the fragment ions present in the spectrum matches only the eight permutations shown in Table 7. Since classical peptide backbone fragmentation does not yield complete sets of b or y ions, the interpretation draws heavily on the internal fragment ions in the spectrum. These internal fragments are a common feature of PSD spectra, and hamper manual attempts at spectral interpretation because of the large number of di-, tri-, and tetrapeptide combinations that one must explore. From a de novo interpretation perspective,[67] these eight sequences are indistinguishable; the only differences are inability to distinguish the order of the first two residues and inability to distinguish I and L at two positions.

We recognized that one relatively obvious use of the list of sequences resulting from de novo MS/MS interpretation as shown in Figure 5 would be to drive complementary sequence homology type searches using algorithms such as FASTA[53] and BLAST.[54] When used to search sequence databases, MS-Tag is generally capable of matching homologous peptides that contain a single AA substitution by implementing the fragment-ion tag concept.[26,68] We believe that matching database sequences with weaker homology would be better facilitated by implementing a de novo interpretation followed by a direct sequence homology search. Independent of our work, this strategy was recognized and implemented by Johnson and Taylor using the program Lutefisk97 for de novo sequence interpretation and CIDentify (a modified version of FASTA) for sequence homology searching.[27] It is not yet clear to us how degenerate the de novo interpretation result pool could be, resulting from MS/MS spectra that yield only partial peptide fragmentation yet still enable reasonable specificity in FASTA[53] and BLAST[54] type searches, particularly when the relatively short peptides typically tractable in MS/MS experiments are used. To begin probing the tolerable level of degeneracy, we modified our text search based program, MS-Edman, to accept a list of possible sequences and perform database searches allowing for a user-designated limited number of AA mismatches. As shown in Table 7, an MS-Edman search using the eight sequences resulting from de novo interpretation of the spectrum in Figure 6 matches only homologous apolipoprotein A-1 sequences from several species when up to two AA mismatches are allowed. With this particular example, when three AA mismatches are allowed, specificity declines significantly with 222 sequences being matched.

If these homology search strategies fail, as one would expect when sequencing an unknown protein, then the de novo interpreted peptide sequences could be used to design degenerate oligonucleotide primers (from two or more peptides sequenced by MS/MS) to be used in PCR-based gene cloning experiments. Furthermore, in high-throughput proteomics work driven by MS/MS interpretation via database searching, de novo sequencing could be used as an overall process quality measurement tool.

(67) Falick, A. M.; Hines, W. M.; Medzihradszky, K. F.; Baldwin, M. A.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 882−893.

(68) Clauser, K. R.; Baker, P. R.; Foulk, R. A.; Fisher, S. J.; Burlingame, A. L. In preparation.

Table 7. De Novo Sequence Interpretation[1] and Database Searches Using MS-Tag with the MS/MS Spectrum in Figure 6

Number of Sequence Mismatches	MS-Tag[1] Matching Strategy	Sequences Matched	Species
0	**Identity mode**[3]	AKPVLEDLR	Cow
1	**Homology mode**[4]	AKPVLEDLR	Cow
		AKP**A**LEDLR	Pig, Macaque, Human
n/a	**Unknome mode**	KAPVIEDIR	n/a
		AKPVIEDLR	
		AKPVLEDIR	
		KAPVIEDLR	
		KAPVLEDIR	
		AKPVLEDLR	
		KAPVLEDLR	
		AKPVIEDIR	
2	**Unknome mode** followed by **MS-Edman**[5]	AKPVLEDLR	Cow
		AKP**A**LEDLR	Pig, Macaque, Human
		ARP**A**LEDLR	Rabbit, Mouse
	Unknome mode followed by **MS-Edman**[5]	AKPVLEDLR	Cow
		AKP**A**LEDLR	Pig, Macaque, Human
		ARP**A**LEDLR	Rabbit, Mouse
		ARP**A**L**Q**DLR	Dog
		+ 218 other sequences	

[1] MS-Tag parameters used in all searches: parent mass tolerance ± 10 ppm, fragment mass tolerance ± 1000 ppm, monisotopic masses, trypsin digest (1 missed cleavge allowed), cysteines modified by acrylamide.
[2] Database: NCBInr release 3/30/97.
[3] Identity mode, all species.
[4] Homology mode, all species, parent mass shift ± 130 Da.
[5] List of sequences mode, all species, 2 and 3 mismatches allowed.

CONCLUSIONS

Improvement in mass measurement accuracy with DE on reflector MALDI-TOF instruments to levels approaching the practical limits of discerning peptide elemental composition has three major implications. First, false positives are nearly eliminated in peptide mass fingerprinting experiments. Homologous proteins of high overall sequence identities (>70%) could be correlated using masses alone if mass measurement accuracy is sufficient to discriminate among peptide elemental compositions. In practice, a single AA substitution search satisfying less strict criteria could be used as a screen to yield a preliminary candidate pool from which MS/MS spectra would serve as the final discriminating tools for identification. This type of strategy is likely to be most useful in studying proteins derived from an organism for which little genome sequence exists but substantial genome sequence from a related organism is available. Second, successful tagging strategies for matching a database sequence which is identical to a single tryptic peptide from the protein being analyzed require little more than an accurate peptide mass and a complete set of immonium ions from PSD + CID. Third, de novo interpretation of MS/MS spectra from peptides with parent mass <1300 Da is possible using a database search algorithm to search a virtual database derived from all permutations of the AA combinations which are not forbidden by the parent mass and immonium ions. The extension of this approach to higher mass peptides may be possible by replacing the fragment-ion tag database search algorithm with a more sophisticated combinatoric approach employing sparse dynamic programming related to the one used in gene recognition.[69] The degenerate peptide sequence, or ambiguous set of possible sequences, resulting from de novo MS/MS interpretation can be used to drive complementary sequence homology type searches using algorithms similar to FASTA[53] and BLAST.[54] In actual practice, homology matching with masses alone and immonium ion tagging with a single peptide represent extreme applications of high-accuracy MS data. More information would be available from a typical combination of MS and MS/MS experiments.

ACKNOWLEDGMENT

Portions of this work were presented in preliminary form at both ABRF '97: Techniques at the Genome/Proteome Interface, a symposium sponsored by the Association of Biomolecular Resource Facilities, Baltimore, MD, Feb 9−12, 1997, and the 44th ASMS Conference on Mass Spectrometry and Allied Topics, Palm Springs, CA, June 1−5, 1997. Financial support for this work was provided by the NIH (Grants NCRR BRTP, RR01614, RR08282, HD30367, and HD22210) and the Ludwig Institute for Cancer Research, London Branch.

Received for review September 22, 1998. Accepted April 9, 1999.

AC9810516

(69) Gelfand, M. S.; Mironov, A. A.; Pevzner, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 9061−9066.