Anal. Chem. **2001,** *73,* 2594−2604

Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry

J. Alex Taylor and Richard S. Johnson*

Immunex Corporation, Seattle, Washington 98101-2936

There are several computer programs that can match peptide tandem mass spectrometry data to their exactly corresponding database sequences, and in most protein identification projects, these programs are utilized in the early stages of data interpretation. However, situations frequently arise where tandem mass spectral data cannot be correlated with any database sequences. In these cases, the unmatched data could be due to peptides derived from novel proteins, allelic or species-derived variants of known proteins, or posttranslational or chemical modifications. Two additional problems are frequently encountered in high-throughput protein identification. First, it is difficult to quickly sift through large amounts of data to identify those spectra that, due to poor signal or contaminants, can be ignored. Second, it is important to find incorrect database matches (false positives). We have chosen to address these difficulties by performing automatic de novo sequencing using a computer program called Lutefisk. Sequence candidates obtained are used as input in a homology-based database search program called CIDentify to identify variants of known proteins. Comparison of database-derived sequences with de novo sequences allows for electronic validation of database matches even if the latter are not completely correct. Modifications to the original Lutefisk program have been implemented to handle data obtained from triple quadrupole, ion trap, and quadrupole/time-of-flight hybrid (Qtof) mass spectrometers. For example, the linearity of mass errors due to temperature-dependent expansion of the flight tube in a Qtof was exploited such that isobaric amino acids (glutamine/lysine and oxidized methionine/phenylalanine) can be differentiated without careful attention to mass calibration.

Protein identifications are most often performed using mass spectrometry. For samples containing only a few proteins, the mass mapping approach is suitable—identifications are made based on comparisons of observed proteolytic peptide masses with peptide masses calculated from a sequence database.[1−7] For more complicated protein mixtures, or for confirmation of results obtained by mass mapping, tandem mass spectrometry (MS/MS) provides sequence-specific fragmentation patterns of individual proteolytic peptides that can be used for sequence database matching. Several database search programs have been written for this purpose.[6,8−11] These programs work most reliably when making exact matches between query peptides and database sequences; however, query and database sequences that are not identical, but homologous, are often overlooked. Likewise, peptides produced by unexpected proteolytic cleavages or modified peptides can also be missed when database matches are made using MS/MS data.

An alternative is to determine de novo sequences from MS/MS data. This approach was demonstrated using peptides of known sequence derived from apolipoprotein B,[12] and shortly thereafter, several thioredoxins of previously unknown sequence from various species were sequenced entirely by MS/MS.[13−16] Although de novo peptide sequencing from MS/MS data is usually done manually, there have been several computer programs devised for automating this procedure. One approach has been to calculate all possible amino acid sequences for a given peptide mass and then determine which sequence best accounts for the ions found in the MS/MS spectrum.[17] A less computationally

* Corresponding author. Telephone: (206) 381-6412. Fax: (206) 621-5440. E-mail: JohnsonR@immunex.com.

(1) 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.
(2) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.
(3) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.
(4) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−345.
(5) Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.
(6) Clauser, K.; Baker, P.; Burlingame, A. *Anal. Chem.* **1999**, *71*, 2871−2882.
(7) Zhang, W.; Chait, B. *Anal. Chem.* **2000**, *72*, 2482−2489.
(8) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(9) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(10) Fenyo, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998−1005.
(11) Perkins, D.; Pappin, D.; Creasy, D.; Cottrell, J. *Electrophoresis* **1999**, *20*, 3551−3567.
(12) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−6237.
(13) Johnson, R. S.; Biemann, K. *Biochemistry* **1987**, *26*, 1209−1214.
(14) Mathews, W. R.; Johnson, R. S.; Cornwell, K. L.; Johnson, T. C.; Buchanan, B. B.; Biemann, K. *J. Biol. Chem.* **1987**, *262*, 7537−7545.
(15) Johnson, R. S.; Mathews, W. R.; Biemann, K.; Hopper, S. *J. Biol. Chem.* **1988**, *263*, 9589−9597.
(16) Johnson, T. C.; Yee, B. C.; Carlson, D. E.; Buchanan, B. B.; Johnson, R. S.; Mathews, W. R.; Biemann, K. *J. Bacteriol.* **1988**, *170*, 2406−2408.
(17) Sakurai, T.; Matsuo, T.; Matsuda, H.; Katakuse, I. *Biomed. Mass Spectrom.* **1984**, *11*, 396−399.

10.1021/ac001196o CCC: $20.00
Published on Web 05/03/2001

intensive approach is to build sequences stepwise beginning with small "subsequences" and, at each step, compare predicted fragmentations to the observed ions.[18−23] Another related approach has been to graph ions onto a "sequence spectrum" and determine the sequence possibilities from this graph.[24−30] More recently, a genetic algorithm[31] has been implemented for de novo peptide sequencing.

We frequently obtain high-quality MS/MS spectra of tryptic peptides for which no obvious database match can be made. For this reason, we chose to develop a computer program that does not rely on a sequence database and instead performs a de novo sequence interpretation of MS/MS data using the graphing approach described above. The output of this program is a short list of candidate sequences, which are used to query sequence databases using a version of the FASTA program[32,33] that has been modified to take into account some of the peculiarities and ambiguities of sequences derived from MS/MS spectra.[29]

Here we describe modifications that have been made to the software, which allow for more effective use of data obtained from three instruments commonly used for proteomic work. Since our original report, our de novo sequencing software for triple quadrupole instruments has undergone further revisions in order to better handle data from ion trap and quadrupole/time-of-flight hybrid (Qtof) mass spectrometers. Compared to the other instruments, data from ion traps exhibit more intense and contiguous series of b-type ions,[34] which requires slightly different rules for interpretation. While Qtof fragmentation patterns are similar to those obtained from triple quadrupoles, Qtof spectra have much higher resolution and mass accuracy, which can be used to aid interpretation. However, we have found that our Qtof mass accuracy varies due to a temperature-dependent expansion and contraction of the flight tube. This difficulty was overcome by exploiting the linearity of errors with respect to mass.

(18) Biemann, K.; Cone, C.; Webster, B. R.; Arsenault, G. P. *J. Am. Chem. Soc.* **1966**, *88*, 5598−5606.
(19) Ishikawa, K.; Niwa, Y. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 373−380.
(20) Siegel, M. M.; Bauman, N. *Biomed. Environ. Mass Spectrom.* **1988**, *15*, 333−343.
(21) Johnson, R. S.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945−957.
(22) Yates, J. R., III; Griffin, P. R.; Hood, L. E. *Computer Aided Interpretation of Low Energy MS/MS Mass Spectra of Peptides*; Academic Press: San Diego, CA, 1991.
(23) Johnson, R. S.; Ericsson, L.; Walsh, K. A. Nashville, TN, 1991; pp 1233−1234.
(24) Bartels, C. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 363−368.
(25) Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V. *Comput. Appl. Biosci.* **1995**, *11*, 427−434.
(26) Fernandez-de-Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Besada, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. *Electrophoresis* **2000**, *21*, 1694−1699.
(27) Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 326−336.
(28) Scarberry, R. E.; Zhang, Z.; Knapp, D. R. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 947−961.
(29) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(30) Dancik, D.; Addona, T.; Clauser, K.; Vath, J.; Pevzner, P. *J. Comput. Biol.* **1999**, *6*, 327−342.
(31) Skilling, J.; Cottrell, J.; Green, B.; Hoyes, J.; Kapp, E.; Langridge, J.; Bordoli, R., Dallas, TX, June 13−17, 1999.
(32) Pearson, W. R. *Methods Enzymol.* **1990**, *183*, 63−98.
(33) Pearson, W. R.; Lipman, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(34) Biemann, K. *Methods Enzymol.* **1990**, *193*, 886−887.

In addition to describing these changes in the computer programs, we illustrate some of the uses we have found for automated de novo peptide sequencing. One application is the validation of database matches, which provides an additional level of assurance that a database hit is not a false positive. Also, nonidentical, homologous sequences can be matched using this approach. Finally, automated de novo sequencing has been useful in identifying chemically modified peptides or peptides derived from nonconsensus proteolytic cleavages.

EXPERIMENTAL SECTION

Lutefisk. Our de novo peptide sequencing algorithm, Lutefisk, is briefly described.[29] The first step is to import either a centroid list of ion masses and intensities or raw profile data. Profile data are smoothed, and a centroided mass value for each ion is extracted by applying a weighted average to local maximums within a mass window that is determined by the instrumental resolution. For triple quadrupole profile data where the fragment ion resolution often varies, the program can automatically determine the resolution for individual spectra. In contrast, ion trap and Qtof spectra typically have a standard fragment ion resolution. Using a variety of filtering techniques, a small set of ions (typically 30−60 ions, depending on the size of the peptide) is selected for sequence analysis.

The next step is to convert the ions into a graph of their corresponding b-type ion masses—the so-called "sequence graph".[24] In the absence of additional information, it is impossible to assign ion types to each ion, so to make the sequence graph each of the observed ions is assumed to be all of the possible ion types. For low-energy CID, this means that an ion is assumed to be both b-type and y-type. For example, an ion of m/z 200.123 in a peptide of molecular weight 1000.542 would have graphed node positions at 200.123 (by first assuming it is a b-type ion) and 802.435 (the corresponding b-type ion mass that would be expected if 200.123 were actually a y-type ion). For triple quadrupole and Qtof MS/MS spectra, ions of m/z greater than the precursor are assumed to be y-type ions only, unless an ion 28 u lower is also present (a potential b/a-type pair). In contrast, ion trap data exhibit a fragmentation pattern where prominent b-type ions are found throughout the mass range; hence, for ion trap data this assumption is not made.

In a previous implementation of this program,[29] the node positions of the graph corresponded to nominal fragment ion masses. In the example above, the nodes would be at 200 and 802, which would be the nominal masses of ions at m/z 200.123 and 802.435. Of course, by converting to nominal masses, the additional information content of high-mass-accuracy data obtainable on Qtof instruments is lost. To avoid throwing away this information, the program was modified so that each node position represents either 0.1, 0.01, or 0.001 u, depending on the error associated with the fragment ion mass measurement. For fragment ion errors greater than 0.2 u, each node position represents 0.1 u, and for fragment errors between 0.02 and 0.2 u, each node position represents 0.01 u. The value assigned to each node position in the sequence graph is determined by an "ion probability",[24] which can be changed by the user and is expected to reflect the relative probabilities of seeing individual ion types in a MS/MS spectrum.

Once the sequence graph has been determined, partial sequences are generated by starting at the N-terminal end of the graph and jumping from node to node in increments corresponding to amino acid monoisotopic masses. Since fragmentation frequently does not occur between the N-terminal amino acids, a provision has been made to allow for as much as a three-amino acid jump from the N-terminus. Additional two-amino acid jumps are allowed throughout the sequence; however, a limit of only one or two such jumps has been set, depending on the peptide mass. Allowing for these two- or three-amino acid jumps compensates for incomplete fragmentation patterns that preclude the determination of a complete sequence yet retains the location and mass of the two- or three-amino acid jumps within the rest of the sequence. Several thousand subsequences can be developed in this manner up until the point where they match the observed peptide mass, at which point they are stored as completed sequences. Tens of thousands of completed sequences are stored for a follow-up scoring, sorting, and ranking.

When the sequence generation has finished, completed sequences are examined to see if they contain long contiguous strings of b-type or y-type ions (for low-energy CID spectra). Sequences that appear to have been derived from alternating b-type and y-type ions are discarded, and this prefiltering often eliminates 90% of the sequence candidates. An additional prefiltering of the remaining sequence list requires that each sequence be able to account for a large fraction of the ion intensity of fragments at masses higher than the precursor ion m/z value. This higher mass region almost always is composed of b-type and y-type ions, and sequences that were derived mostly from the low-mass fragments (those less than the precursor m/z) are discarded. This second prefiltering of the completed sequences often eliminates another 90% of the remaining sequences. At this point, the sequences are ranked according to the percentage of fragment ion intensity that can be accounted as b-type, a-type, y-type, water, or ammonia neutral loss from the b-, a-, or y-type ions, internal cleavage ions lacking both the N- and C-termini of the peptide, or immonium ions. The current version also considers losses of CH_3SOH (63.998 u) from b-type and y-type ions containing oxidized methionine.[35] When a calculated ion mass is near the error tolerance cutoff compared to the observed fragment mass, the contribution of the ion intensity to the summed intensity score is attenuated depending on the magnitude of the error. Also, certain ion types will contribute less to the summed intensity score; for example, b- and y-type ions are weighted more heavily compared to internal fragment ions. Program output consists of a list of five or fewer candidate sequences.

CIDentify. The program CIDentify is a modified version of the FASTA homology-based sequence database search program[32,33] and has been described previously.[29] In short, it performs a FASTA homology search of a sequence database and has been modified to allow for multiple query searches in order to accommodate the multiple sequence candidates produced by Lutefisk. In addition, it can deal with the peculiarities and ambiguities of sequences obtained by de novo sequence determinations from MS/MS data. For example, unsequenced dipeptide masses present in a sequence obtained from Lutefisk can be utilized in a database search. Dipeptides in a query that have the same mass as a database dipeptide are recognized, and CIDentify knows that some amino acids have the same mass as certain dipeptides (e.g., Asn equals the sum of two Gly residues). CIDentify can perform nucleotide database searches via six-frame translation.

CIDentify Results Compilier. The output from several peptides can be combined using a program called CIDentify Result Compiler.[36] By examining the results for several peptides, homologies are highlighted.

Data Acquisition. Triple quadrupole and ion trap data were obtained using a Finnigan (San Jose, CA) TSQ700 or LCQ, respectively. Both instruments were equipped with a home-built nanospray platform using electrospray ionization voltages of 600−700 V. Precursor ion resolution for the triple quadrupole was sufficient to pass a 4−5 u wide window into the collision region, and the product ion resolution was ~500 (peaks were often 2 u wide). Argon collision gas was supplied at a pressure of 4 mTorr. Operation of the Finnigan LCQ utilized standard factory settings; maximum trap times were set to 1 s, and approximately 50 scans (each composed of three miniscans) were signal averaged in profile mode. Quadrupole/time-of-flight hybrid Qtof data were obtained using the Micromass (Beverly, MA) Qtof 1. A microspray source for capillary LC/MS was devised using uncoated New Objective tips (Cambridge, MA), where the electrospray voltage was applied with a liquid metal junction via a platinum wire inserted into a tee. Most peptides were obtained via in-gel tryptic digestion using established protocols.[37,38] For nanospray ionization, extracts were desalted on disposable microcolumns prepared using Poros 50R2 resin packed in a Eppendorf (Brinkmann Instruments, Westbury, NY) Geloader pipet tip.[39] Desalted peptides were eluted directly into nanospray tips obtained from either Protana (Odense, Denmark) or New Objective (Cambridge, MA).

Program Availability. Lutefisk is written in ANSI-compliant C and has been implemented on several platforms—Macintosh, Windows NT, and different versions of UNIX, including Linux. Source code, and executables for Macintosh and Windows NT, can be downloaded (http://www.immunex.com/researcher/lutefisk/). CIDentify source code can be downloaded from the University of Virginia FASTA site (ftp://ftp.virginia.edu/pub/fasta).

RESULTS AND DISCUSSION

Enhancements to Lutefisk. Several improvements have been made to Lutefisk that increase the likelihood of deriving a correct sequence. For many tryptic peptide spectra there are a series of y-type ions at masses greater than the precursor ion that clearly delineate partial sequences—the so-called "peptide sequence tags".[9] By using a peptide sequence tag as the starting point, Lutefisk can quickly determine complete sequence candidates that are more likely to be correct. By incorporating these automatically

(35) Swiderek, K.; Davis, M.; Lee, T. *Electrophoresis* **1998**, *19*, 989−997.

(36) Johnson, R.; Taylor, J. In *Methods in Molecular Biology: Mass Spectrometry of Proteins and Peptides*; Chapman, J., Ed.; Humana Press: Totawa, NJ, 2000; Vol. 146, pp 41−62.

(37) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.

(38) Hellman, U.; Wernstedt, C.; Gonez, J.; Heldin, C.-H. *Anal. Biochem.* **1995**, *224*, 451−455.

(39) Erdjument-Bromage, H.; Lui, M.; Lacomis, L.; Grewal, A.; Annan, R.; McNulty, D.; Carr, S.; Tempst, P. *J. Chromatogr., A* **1998**, *826*, 167−181.

Table 1. Automated de Novo Sequence Determinations Using Ion Trap Data

no.	correct sequences[a]	sequence deduced by Lutefisk[b]	rank[c]	correct[d]	incorrect[e]
		Correct Lutefisk Sequences			
1	ANRPFLVFIR	[185.0]RPFLVFLR	1/1	database	database
2	FRIEDGFSLK	FRLEDGFSLK	2/2	database	de novo
3	LQPLDFKENAEQSR	[241.2]PLDFKENAEKSR	1/1	database	de novo
4	TSDQIHFFFAK	TSDKLHFFFAK	1/5	database	de novo
5	CCTESLVNR	CCTESLVNR	1/4	database	de novo
6	LVNELTEFAK	LVNELTE[218.1]K	1/2	database	de novo
7	TVMENFVAFVDK	[200.0]MENFVAFVDK	1/3	database	de novo
8	YICDNQDTISSK	YLCDNKDTLSSK	3/5	database	de novo
9	YLYEIAR	YLYELAR	1/1	database	de novo
10	VLITTDLLAR	[212.1]LTTDLLAR	1/2	database	de novo
11	HTLNQIDSVK	HTLNKLDWK	2/4	database	de novo
12	AGALNSNDAFVLK	KALNSNDAFVLK	1/1	database	de novo
13	LYEIGAGTSEVR	LYELGKTSEVR	1/2	database	de novo
14	ETLEEVFEK	[230.0]LEEVFEK	1/2	database	de novo
15	DGFPSGAPALNTK	[172.0]FPSGAP[184.1]NTK	3/5	database	de novo
16	LLQTSNITK	LLKTS[227.1]TK	2/5	database	de novo
17	NAAQFLLSTNDK	[185.0]AKFLLSTNDK	3/5	database	de novo
18	VVIFQQEQENK	[198.1]LFKKEKENK	1/1	database	de novo
		Incorrect Lutefisk Sequences			
19	AFLEVNEEGSEAAASTAVVIAGR	[262.1][198.1]VNEESGEAAASTAVVLKR	5/5	database	de novo
20	ATEDEGSEQKIPEATNR	[274.0]A[200.0]GSEKKLPEA[215.0]R	2/5	database	de novo
21	DDLYVSDAFHK	[229.0]NYVS[198.1]S[185.0]K	1/5	database	de novo
22	AEFVEVTK	AEFGA[199.1]TK	4/5	database	de novo
23	DDPHACYSTVFDK	[230.0]PHACYG[230.0]FDK	2/5	database	de novo
24	ETYGDMADCCEK	[230.0]YGDMAD[210.2][211.1]R	1/2	database	de novo
25	EYEATLEECCAK	Y[214.1]CEELTAFKK	5/5	database	de novo
26	ATQALVLAPTR	ATKALVLATPR	1/5	database	de novo
27	GYDVIAQAQSGTGK	[335.2]VLAKAGASGTGK	4/5	database	de novo
28	QDGQFSVLFTK	[243.1][185.0]FSVLFTK	1/2	database	de novo
29	GFSVVADTPELQR	[333.2]GVADT[226.1]LKR	2/5	database	de novo
30	LEYFGDH	LEY[174.0]PGK	1/2	de novo	de novo
31	LTDFGFCAQITPEQSK	[361.2][226.1]TLAGACFGDKSK	1/5	database	de novo
32	LFEEPEDPSNR	LFEEPED[212.1]NK	1/4	database	de novo

[a] The underlined letters indicate regions of the correct sequences that were delineated by contiguous sequence-specific fragment ions in the spectrum. [b] The sequence derived by Lutefisk that is most correct is shown. The mass accuracy of the ion trap was not sufficient to differentiate Gln from Lys, and "K" indicates either of these amino acids. Leu and Ile are not distinguishable and are denoted as "L". Numbers in brackets denote summed amino acid residue masses. [c] The ranking of the correct sequence in the output is denoted in the numerator; the denominator indicates the total number of sequence candidates in the output. [d] Indicates whether a correct database-derived sequence or a de novo sequence best accounts for the CID data. [e] Indicates whether an incorrect database-derived sequence or a de novo sequence best accounts for the CID data. The incorrect database-derived sequence that was chosen for comparison was always the second ranked database match obtained using the database search program Mascot.

determined sequence tags into the "sequence graph", the program is forced to derive completed sequence candidates that only contain these partial sequences. As a consequence, the processing time drops, and the quality of the candidate sequences increases. Since ion trap data contain both b- and y-type ions in the high-m/z region (see below), the use of automatically determined "peptide sequence tags" is not recommended for these spectra.

Lutefisk was originally designed only for interpretation of triple quadrupole data from tryptic peptides; however, we find that ion trap spectra exhibit a slightly different fragmentation pattern. Triple quadrupole and Qtof data usually contain a contiguous series of y-type ions throughout the mass range and b-type ions that are most abundant at lower masses and often absent at higher masses. Presumably, this observation is due to the multiple collisions occurring in a quadrupole collision cell; if b-type ions are less stable than y-type ions, then additional collisions would reduce the abundance of higher mass b-type ions compared to y-type ions. In contrast, fragmentation in an ion trap instrument will cause the fragment ions to go out of resonance, thereby halting or reducing further cleavage of the b-type ions. To account for this difference, Lutefisk assumes each ion is both b-type and y-type throughout the mass range when building the sequence graph for ion trap data.

To test the results for ion trap data, 32 MS/MS spectra were sequenced using Lutefisk (Table 1). In most cases there were incomplete sequence-specific fragmentations (either b-type or y-type ions) such that sections of most sequences could not be delineated; regions of the peptide that could be sequenced are underlined in column 2. Despite the incomplete fragmentation, Lutefisk could correctly sequence 18 of the peptides. For example, for peptide 15 there were no b-type or y-type fragment ions indicating cleavage between Asp(1)−Gly(2) or Ala(9)−Leu(10). Nevertheless, Lutefisk determined a sequence containing summed residue masses of 172.0 (Asp + Gly) and 184.1 (Ala + Leu) at the right positions in an otherwise correct sequence. There were 14 other peptides that were not correctly sequenced. However, by comparing the regions delineated by sequence-specific ions (Table 1, column 2) with the sequences determined by Lutefisk (Table 1, column 3), one can see that the program did a reasonable job given the quality of the data. The columns labeled "correct" and "incorrect" in Tables 1−3 are discussed later. Lutefisk performed fairly well using ion trap data—it correctly sequenced

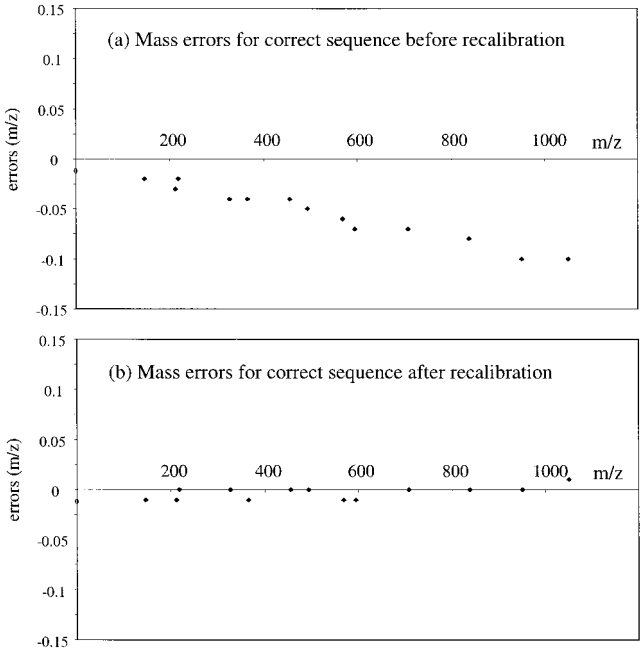


Figure 1. Effect of recalibrating CID data using b-type and y-type ion masses calculated from the correct sequence, LVNELTEFAK. (a) The b-type and y-type ion masses calculated from the correct sequence were compared to the observed fragment masses, and the errors are plotted against the fragment ion masses. Although at the higher mass values errors approach −0.1 u, the plot of the errors is almost linear across the mass range. (b) Using a least-squares fit to the data in panel a, a linear calibration correction was applied to the observed mass values. The errors following such recalibration are better than 0.02 u throughout the mass range.

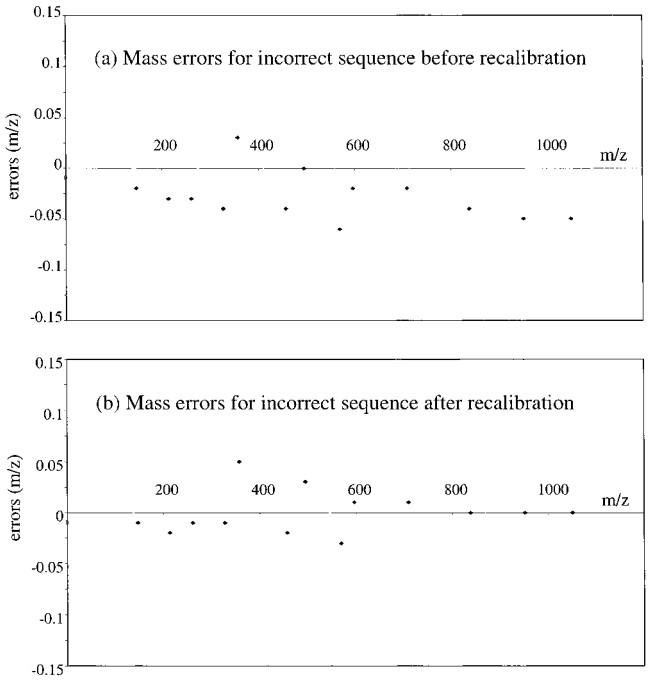


Figure 2. Effect of recalibrating CID data using b-type and y-type ion masses calculated from an incorrect sequence, LVNELTHPLK. (a) The b-type and y-type ion masses calculated from this incorrect sequence were compared to the observed fragment masses, and the errors are plotted against the fragment ion masses. Although the errors for this incorrect sequence are less than those derived from the correct sequence (Figure 1a), they exhibit a greater degree of scatter. (b) Using a least-squares fit to the data in panel a, a linear calibration correction was applied to the observed mass values. The errors resulting from recalibration using calculated b-type and y-type ions from this incorrect sequence are shown to be greater than what can be achieved from the correct sequence (Figure 1b).

most of the peptides, and for less well-defined data it often derived a partially correct sequence.

The high-mass accuracy of the Qtof mass spectrometer should, in principle, provide data that can be sequenced more reliably than data obtained from lower mass accuracy instruments; however, one of the difficulties in the use of this instrument has been the dependence of the Qtof mass accuracy on temperature. For our instrument, mass accuracy varies by 50 ppm/°C. In our laboratory, the temperature has varied by as much as 4 °C over the course of a day, which translates to an error of 0.2 u at m/z 1000. Since most of our samples are derived from in-gel tryptic digestion, we can apply a linear correction for each LC/MS/MS file using the fragment ions obtained from the trypsin autolysis peptides. Despite this calibration correction, the mass accuracy could continue to drift slightly over the course of an LC/MS/MS run; also we find that some users may on occasion neglect to perform this correction. However, although the absolute errors can be as much ±0.2 u, they are linear with respect to the mass-to-charge ratio. For example, a plot of the errors for b-type and y-type ions versus m/z for the peptide LVNELTEFAK is quite linear (Figure 1a). In this example, the highest absolute error is around −0.1 u. Using the least-squares method, the slope of a line can be determined and used to correct the fragment ion m/z values (Figure 1b). After applying this correction, the maximum error is ∼0.02 u.

The linearity of the Qtof errors is exploited in Lutefisk by using a wide fragment ion tolerance to generate numerous candidate sequences, each of which is subsequently reevaluated using a tighter tolerance following a mass correction. Hundreds of candidate sequences are determined using a loose tolerance of ±0.25 u. As above, a linear correction using the least-squares method is determined from the differences between b-type or y-type ions calculated for each candidate sequence and the observed ion m/z values. The correction is applied to the MS/MS data, and the sequence is reevaluated using a tighter error tolerance of ±0.02 u. A new correction is determined for each candidate sequence. For example, in addition to the correct sequence LVNELTEFAK (Figure 1), Lutefisk generates many other incorrect candidate sequences (e.g., LVNELTHPLK (Figure 2)). Before mass correction, the absolute errors for the incorrect sequence (Figure 2a) are less than those observed for the correct sequence (Figure 1a); however, after mass correction, many of the ions for the incorrect sequence are in error by more than 0.02 u. Hence, by determining a mass correction for each candidate sequence and reevaluating them with a tighter tolerance, correct sequences are more readily differentiated from incorrect ones. The same concept could also be applied in the evaluation of sequences obtained from database matches.

Lutefisk correctly determined the sequences of 50 peptides using Qtof data (Table 2). In these examples, our definition of a correctly determined sequence is much more stringent than for the lower mass accuracy ion trap data (Table 1). Specifically, Gln and Lys (residue masses of 128.059 versus 128.095 u, respectively), and oxidized Met and Phe (residue masses of 147.035 versus

Table 2. Correct Automated de Novo Sequence Determinations Using Qtof Data

no.	correct sequences[a]	sequences deduced by Lutefisk[b]	rank[c]	correct[d]	incorrect[e]
1	LALDLEIATYR	[184.12]LDLELATYR	1/2	database	de novo
2	QFSSSYLSR	A[204.09]SSSYLSR	1/1	database	de novo
3	IPDGFAGAQGGITFR	LPDGFAGAQG[170.11]TFR	1/5	NA	de novo
4	IPIGFAGAQGGFDTR	LPLGFAGAQG[204.09]DTR	1/5	NA	de novo
5	FALPQYLK	FALPQYLK	1/2	database	de novo
6	ALNEINQFYQK	[184.12]NELNQFYQK	1/5	database	de novo
7	YLGYLEQLLR	YLGYLEQLLR	1/2	database	de novo
8	MPCTEDYLSLINR	MPCTEDYLSLLLNR	1/2	database	de novo
9	QNCDQFEK	QNCDQFEK	1/2	database	de novo
10	CCTESLVNR	CCTESLVNR	1/1	database	de novo
11	CCAADDKEACFAVEGPK 3+	CCAADDKEACFAVEGPK	1/5	database	de novo
12	CCAADDKEACFAVEGPK 2+	CCAADDQEACFAVEGPK	1/5	database	de novo
13	VLDALDSIK	VLDALDSLK	1/2	database	de novo
14	ELVISLIVESK	[242.12]VLSLLVESK	1/2	NA	de novo
15	VLDELTLTK	VLDELTLTK	1/1	database	de novo
16	AEFVEVTK	AEFVEVTK	1/1	database	de novo
17	LSSPATLNSR	LSSPATLNSR	1/1	database	de novo
18	ELVISISDEADK	[242.12]VLSLSDEADK	1/5	NA	de novo
19	SLVNLGGSK	SLVNLGGSK	1/2	database	de novo
21	QEYEQLIAK	[257.10]YEQLLAK	1/5	database	de novo
22	QVLDNLTmEK	[227.13]LDNLTmEK	1/5	database	de novo
23	IEISELNR	IELSELNR	1/1	database	de novo
24	YEELQVTVGR	[EY]ELQVTVGR	1/2	database	de novo
25	YEELQITAGR	YEELQLTAGR	1/2	database	de novo
26	TNAENEFVTIK	[215.09]AENEFVTLK	1/2	database	de novo
27	TLLDIDNTR	[214.13]LDLDNTR	1/2	database	de novo
28	LNDLEDALQQAK	LNDLEDALQQAK	2/5	database	de novo
29	SLDNQFASFIDK	[200.11]DNQFASFLDK	1/5	database	de novo
30	SLDLDSIIAEVK	[200.11]DLDSLLAEVK	1/3	database	de novo
31	DAFLGSFLYEYSR	[186.06]FLGSFLYEYSR	3/5	database	de novo
32	LGEYGFQNAILVR	LGEYGFQNALLVR	1/5	database	de novo
33	SVLGDVGITEVFSDR	[186.10]LGDVGLTEVFSDR	1/5	database	de novo
34	TVMENFVAFVDK	[200.11]MENFVAFVDK	1/5	database	de novo
35	VLDPNTVFALVNYISFK	[LV]D[NP]TVFALVNYLSFK	1/5	database	de novo
36	LVNELTEFAK	LVNELTEFAK	1/2	database	de novo
37	TVmENFVAFVDK	[200.11]mENFVAFVDK	1/5	database	de novo
38	QDGQFSVLFTK	[243.09]GQFSVLFTK	1/5	database	de novo
39	LVDTFLEDVK	LVDTFLEDVK	1/2	database	de novo
40	HLVDEPQNLIK	HLVDE[225.11]NLLK	2/2	database	de novo
41	YLYEIAR	YLYELAR	1/1	database	de novo
42	EYEATLEECCAK	[EY]EATLEEC[AC]K	1/1	database	de novo
43	YICDNQDTISSK	YLCDNQDTLSSK	1/2	database	de novo
44	DDPHACYSTVFDK 3+	[230.06]PHACYSTVFDK	4/5	database	de novo
45	DFPIANGER	DFPLAN[186.07]R	2/2	database	de novo
46	GITWGEETLMEYLENPK	[170.11]TWGEETLMEYLENPK	2/5	database	de novo
47	FFVAPFPEVFGK	FFVAPF[EP]VFGK	1/5	database	de novo
48	VTPEIPAGLPSPR	VTPELPA[170.11][184.08]PR	3/5	database	de novo
49	AAQDSFAAGWGVMVSHR 3+	AAQDS[218.11]AGWGVMV[HS]R	2/5	database	de novo
50	FKDLGEEHFK 3+	KFDLGEEHFK	1/5	database	de novo

[a] The underlined letters indicate regions of the correct sequences that were delineated by contiguous sequence-specific fragment ions. MS/MS spectra acquired from triply charged precursors are denoted by "3+" following the sequence, and all others were obtained from doubly charged precursors. [b] The sequence derived by Lutefisk that is most correct is shown. Numbers in brackets denote dipeptide residue masses. In cases where the dipeptide mass matched a specific pair of amino acids, the program lists the pair in brackets. L denotes both Leu and Ile; m denotes oxidized Met; given the mass accuracy of the Qtof, Q and K are distinguishable. [c] The ranking of the correct sequence in the output is denoted in the numerator; the denominator indicates the total number of sequence candidates in the output. [d] Indicates whether a correct database-derived sequence or a de novo sequence best accounts for the CID data. NA indicates peptides not present in the sequence database. [e] Indicates whether an incorrect database-derived sequence or a de novo sequence best accounts for the CID data. The incorrect database-derived sequence that was chosen for comparison was always the second ranked database match obtained using the database search program Mascot.

147.068 u, respectively) are correctly differentiated. In addition, Lutefisk was able to distinguish between Trp and dipeptides of residue masses nearly equal to Trp (peptides 31, 33, and 45). The numbers in brackets indicate summed amino acid residue masses that cannot be sequenced; however, in a few cases, the mass accuracy is sufficient to correctly identify a specific pair of amino acids (peptides 24, 35, 42, 47, and 49). In most cases, the correct sequence is ranked first out of five or fewer candidates. Data were reduced to 20−50 fragment ions prior to sequencing, and the calculations for each spectrum were completed within a few seconds on a 333-MHz Macintosh G3 computer.

Table 3 shows the results from 13 peptides where Lutefisk did not correctly determine sequences using Qtof data. For peptides 1−4, the program confused Gln for Gly−Ala (1), oxidized Met for Phe (2 and 4), and Trp for Gly−Glu (3). The data for peptides 5−11 lacked fragment ions delineating three or more contiguous amino acids in the sequence. Although the sequences derived by Lutefisk were not completely correct, the regions that

Table 3. Incorrect Automated de Novo Sequence Determinations Using Qtof Data

no.	correct sequences[a]	sequences deduced by Lutefisk[b]	correct[c]	incorrect[d]
		Confused Q and GA, Oxidized M (′m′) and F, or W and GE		
1	SIVPSGASTGVHEALEMR 3+	AEVPSQSTGVHEALEMR	database	de novo
2	DDPHACYSTVFDK	[327.11]HACY[188.08]VmDK	database	de novo
3	DFPIADGER	[262.10]PLADWR	database	de novo
4	SLNNQFASFIDK	[200.11]NNQFASmLDK	database	de novo
		CID Data with Extensive Regions Lacking Sequence Fragmentations		
5	NAVPITPTLNR	[185.08]VP[HM]V[300.12]K	database	de novo
6	EPISVSSQQmLK	[200.11][EP]VS[242.12]VDFK	database	de novo
7	EPQVYVLAPPQEELSK	[225.11]EVYV[NP]AMVY[202.06]SK	database	de novo
8	TGPNLHGLF	TG[314.14][MW]F	NA	de novo
9	NQELAYFYPELFR	[242.10]ELAYFY[HV][HL]R	NA	de novo
10	ETYGDMADCCEK	[393.12]GDMADCCEK	database	de novo
11	AVVQDPALKPLALVYGEATSR 3+	[170.11]VQ[310.13][214.13][LP]ALVYGE[259.10]R	database	de novo
		Other		
12	HPEYAVSVLLR 3+	[284.16][198.10]DVSVLLR	database	de novo
13	HPYFYAPELLYYANK 3+	[243.10][230.09]A[234.10]PELLYYANK	database	de novo

[a] The underlined letters indicate regions of the sequence that were delineated by contiguous sequence-specific fragment ions. MS/MS spectra acquired from triply charged precursors are denoted by "3+" following the sequence, and all others were obtained from doubly charged precursors. [b] Numbers in brackets denote dipeptide residue masses. In cases where the dipeptide mass matched a specific pair of amino acids, the program lists the pair in brackets. L denotes both Leu and Ile; m denotes oxidized Met; given the mass accuracy of the Qtof, Q and K are distinguishable. [c] Indicates whether a correct database-derived sequence or a de novo sequence best accounts for the CID data. NA indicates peptides not present in the sequence database. [d] Indicates whether an incorrect database-derived sequence or a de novo sequence best accounts for the CID data. The incorrect database-derived sequence that was chosen for comparison was always the second ranked database match obtained using the database search program Mascot.

could be sequenced were often correct. Peptides 12−13 exhibited sequence-specific fragment ions delineating the entire sequences, except for dipeptides at the N-terminus. Lutefisk did not correctly sequence these peptides; however, much of the C-terminal regions were right.

Although automated de novo sequencing is possible, one might reasonably question its utility, given the expansion of sequence databases and the ease with which they can be searched. At present, there are only a limited number of completed genome sequences of variable quality, and mass spectrometric identification of proteins obtained from species for which there is incomplete data will require de novo sequencing. In addition, we have found other uses for automated de novo sequencing, which are detailed below.

Validating Database Matches. Is there any point in performing automated de novo sequencing in cases where a database match can be made? Ideally each database hit should be independently validated; however, the volume of data generated in LC/MS/MS experiments precludes the possibility of manually checking every match. The obvious and simplest validation is to check if several different peptides are found to match the same protein, but in some cases, there may be only a few or a single peptide from a particular protein. In these cases, it may be possible to electronically validate a database match by comparing the database sequences with automated de novo sequence interpretations.

It is possible to combine the multitude of candidate de novo sequences generated by Lutefisk with the relatively few sequences identified using database search programs and then allow Lutefisk to score, rank, and compare the combined list of sequences (Figure 3). From such a comparison, Lutefisk can assess whether a database match is correct regardless of whether the automated de novo sequencing results were correct. The de novo sequencing algorithm is free to use any possible sequence to account for the fragmentation data, whereas the range of possible sequences that can be obtained from a database are more constrained. Hence, the best de novo sequences (including incorrect ones) will usually account for the observed fragment ions better than incorrect database-derived sequences. On the other hand, correct database sequences will be as good or better than the best de novo

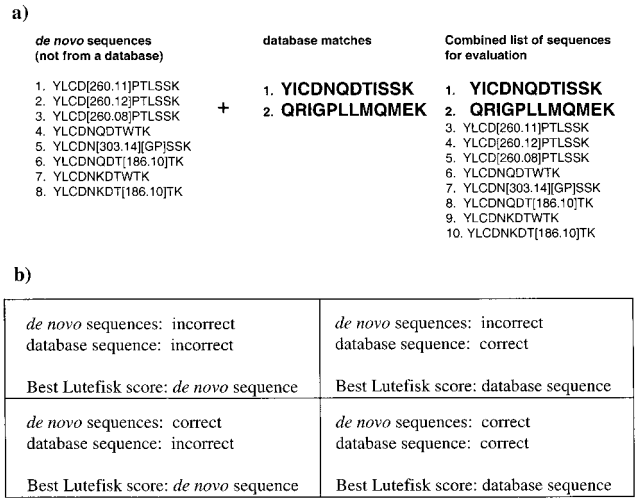


Figure 3. Using automated de novo sequencing to validate database matches. (a) The numerous sequence candidates that are generated by Lutefisk are combined with the relatively few database-derived sequences. The combined list of sequences is then scored and ranked in the final stage of Lutefisk processing. (b) Since the de novo sequence candidates and the database-derived sequences can either be correct or incorrect, there are four possible outcomes when both results are considered together. Even if the de novo sequence candidates are wrong, Lutefisk will generally assign a higher ranking to de novo sequences compared to incorrect sequences obtained from a database search. In contrast, correct database sequences will generally be ranked as high or higher than de novo sequence candidates.

sequences. This leaves four possibilities, and the predicted outcomes are depicted in Figure 3b.

By combining correct de novo and correct database-derived sequences prior to final evaluation by Lutefisk, the prediction (Figure 3b) is that correct database-derived sequences will account for the MS/MS data as well or better than the de novo sequences. For peptides 1−18 in Table 1, where Lutefisk correctly determined a de novo sequence from ion trap data, the designation "database" in the column labeled "correct" signifies that the database-derived sequence matched the MS/MS data as well or better than any of the de novo sequences. Likewise, the "correct" column in Table 2 shows the same results for data obtained from a Qtof.

By combining incorrect de novo sequences with correct database sequences prior to final evaluation by Lutefisk, the prediction (Figure 3b) is that the correct database-derived sequences will account for the MS/MS data better than incorrect de novo sequences. For peptides 19−32 in Table 1, where Lutefisk did not provide correct sequences from ion trap data, the correct database-derived sequences were, with a single exception, ranked highest. For the incorrect de novo sequences obtained from Qtof data (Table 3), the correct database-derived sequences also were ranked highest.

By combining correct de novo and incorrect database-derived sequences prior to final evaluation by Lutefisk, the prediction is that the correct de novo sequences will account for the MS/MS data better than incorrect database sequences. To test this prediction, the Lutefisk candidate sequences were compared against the second-ranked (and incorrect) sequences obtained using the database search program Mascot.[11] Of the first 18 peptides, such a comparison (Table 1, column labeled "incorrect") favored the de novo sequences in all but one case (peptide 1). Similar results were obtained from the Qtof data (Table 2 column labeled "incorrect"), where a comparison of the incorrect second-ranked sequence obtained by Mascot favored the correct de novo sequences in all 50 cases. Peptides 3, 4, 14, and 18 are not found in any databases, so the top-ranked (and incorrect) sequence obtained by Mascot was used in the comparison.

By comparing incorrect de novo with incorrect database-derived sequences prior to final evaluation by Lutefisk, the prediction is that the incorrect de novo sequences will account for the MS/MS data better than incorrect database sequences. To test this prediction, the Lutefisk candidate sequences were compared against the second-ranked (and incorrect) sequences obtained using the database search program Mascot. Lutefisk incorrectly sequenced peptides 19−32 (Table 1), yet by comparing these candidates with the incorrect second-ranked database-derived sequences, the de novo sequences were better able to account for the MS/MS data obtained from an ion trap (Table 1, column labeled "incorrect"). Similar results were obtained for Qtof data, where the de novo sequences accounted for the MS/MS data better than the second-ranked database-derived sequences obtained using Mascot (Table 3, column labeled "incorrect"). Peptides 8 and 9 were derived from nontryptic cleavages and could not be correctly matched in a database search, so the top-ranked Mascot peptide was compared to the de novo sequences. In all of these cases, the incorrect de novo sequences matched the MS/MS data better than the incorrect database-derived sequences.

Table 4. Sequencing Results for A39R Receptor Peptides

no.	sequence	rank[a]	plexin[b]
1	LLLPYR	1/2	no
2	SPTTTALCLFR	2/4	no
3	TVLFLGTGDGQLLK	0/10	no
4	EETPVFYK	1/2	no
5	NIYIYLTAGK	1/6	no
6	TTVTMVGSFSPR	1/6	no
7	ECPACVETGCAQCK	2/10	no
8	TNQALQVFYIK	1/9	no
9	FSLPSSR	2/4	no
10	VQDTYLDCGTLQYR	1/1	no
11	NQDLTTILCK	1/2	no
12	DGFAELQMDK	1/3	no
13	LLTNWMSVCLSGFLR	2/6	yes
14	VAIHSVLEK	1/1	no
15	SIWSLPNSR	4/4	no
16	YFFDFLDAQAENK	1/6	yes
17	FWVNILK	0/1	yes
18	NPQFVFDIK	1/1	yes
19	YFDEILNK	1/2	no

[a] The numerator indicates the rank of the correct sequence, and the denominator contains the total number of sequence candidates produced. [b] CIDentify identified a plexin family member as a homologous nonidentical match.

In conclusion, if de novo sequences (correct or incorrect) account for the CID data better than a database-derived sequence, then the database sequence is likely to be a false positive. If the database-derived sequence accounts for the CID data better than or equal to the de novo sequences, then it is likely to be a real match. Hence, it is possible to use automated de novo sequencing to assist in electronically validating sequence database matches.

Finding Homologous Nonidentical Proteins. For proteins not in any sequence database, can tandem mass spectrometry be used to identify related proteins that are in a database? Automated de novo sequencing followed by homology-based database searches can be used to identify proteins that are known in other species. For example, a kinase isolated from rabbit lungs was shown to be homologous to human and mouse protein kinase C and was therefore presumed to be the rabbit homologue.[36] We have made similar identifications for certain bovine serum proteins (e.g., c-reactive protein (data not shown)), where database searches yield no matches, yet further analysis demonstrated that the proteins were most likely known proteins of bovine origin.

A greater challenge is to determine whether a completely unknown protein contains sequence motifs that place it within a protein family. The vaccinia viral protein, A39R, was used to affinity purify an A39R receptor from a human B cell line.[40] A protein of molecular weight 220 000 was isolated as a band on a SDS polyacrylamide gel electrophoresis gel, excised, and subjected to a standard in-gel tryptic digestion. The tryptic peptides were analyzed using a triple quadrupole instrument equipped with a nanospray source, and MS/MS spectra were obtained for 19 peptides (Table 4). A protein sequence database search using the program Sequest (beta version 22)[8] revealed no identical matches.

(40) Comeau, M.; Johnson, R.; DuBose, R.; Petersen, M.; Gearing, P.; VandenBos, T.; Park, L.; Farrah, T.; Buller, R.; Cohen, J.; Strockbine, L.; Rauch, C.; Spriggs, M. *Immunity* **1998**, *8*, 473−482.

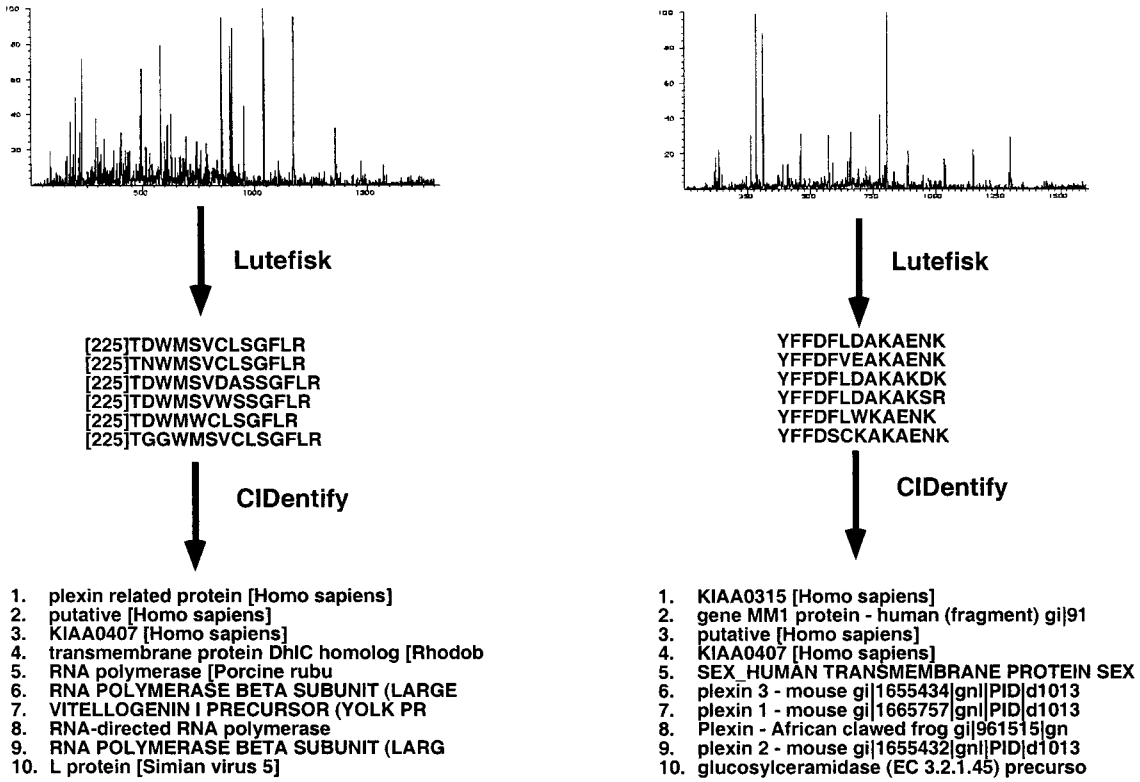


Figure 4. Homologous matches to the plexin protein family. Two MS/MS spectra of tryptic peptides from VESPR are shown. Following automated de novo sequencing using the program Lutefisk, six candidate sequences were determined for each spectrum. The numbers in brackets are the summed residue masses of two unsequenced amino acids. Each set of candidate sequences was used as input for a homology-based database search using a modified version of FASTA called CIDentify. The proteins in the database that provided the most homologous matches are listed at the bottom; many of these proteins are members of the plexin family.

The Lutefisk output files for the tryptic peptides were searched against a nonredundant protein database using CIDentify, and in a few cases, homologous matches were made to the plexin family of proteins (Figure 4 and Table 4). The CIDentify Results Compiler was used to locate database entries that appeared in multiple CIDentify homology search results where it was found that the top entries were members of the plexin family (Table 4). Hence, even before the complete cDNA sequence was known, it was clear from the MS/MS spectra that this protein was also a likely plexin family member.

A database search revealed identical matches for two human peptides to the same human EST sequence, and this information was eventually used to clone the full-length sequence.[40] In addition to this sequence pair, another identical match of a human peptide was made to a mouse EST, and by using Lutefisk in conjunction with CIDentify, a second nonidentical human peptide was matched to the same mouse EST (Figure 5). The only sequence difference between the human peptides and the mouse translated sequence (Figure 5a) is that the human peptide has a C-terminal Arg, whereas the mouse sequence from the EST database has Leu in the corresponding position (Figure 5c). The single amino acid difference (Figure 5a) not only shifts the peptide mass but also eliminates a predicted tryptic cleavage site, therefore precluding the possibility of using software designed for making identical database matches. It is important to point out the importance of having two peptides match the same EST sequence, since it is quite easy to make random matches using single short queries in such large nucleotide sequence databases.

Identifying Chemically Modified Peptides and Peptides Derived from Nonconsensus Protease Cleavages. Technically related to the difficulties associated with finding homologous nonidentical matches to database-derived sequences is the problem of identifying chemically modified database-derived sequences. In both situations, the mass calculated from the database does not match what is observed, thereby making an identification considerably more difficult. Database search programs can be informed of likely modifications (e.g., oxidized methionine and acrylamide adducts); however, by allowing too many possible modifications one increases the number of database sequences with the correct mass, thereby increasing the possibility of obtaining a false positive match. A similar problem exists for de novo sequencing, since sequences are typically built using only the standard amino acids. As with database searching, Lutefisk can be informed of any suspected modified residues, but doing this will increase the number of sequence possibilities. Nevertheless, we have on occasion made identifications of modified peptides on the basis of automated de novo sequencing results. Since Lutefisk routinely generates sequence candidates containing an unspecified mass at the N-terminus, it can easily determine sequences of peptides containing unexpected N-terminal modifications. An example of this was shown earlier,[41] where a peptide with an N-terminal carbamidomethylated cysteine had cyclized and lost ammonia. The sequence determined by Lutefisk exactly matched the database sequence; however, the N-terminal dipep-

(41) Johnson, R. In *Proteome Research: Mass Spectrometry*; James, P., Ed.; Springer-Verlag: Berlin, 2001.

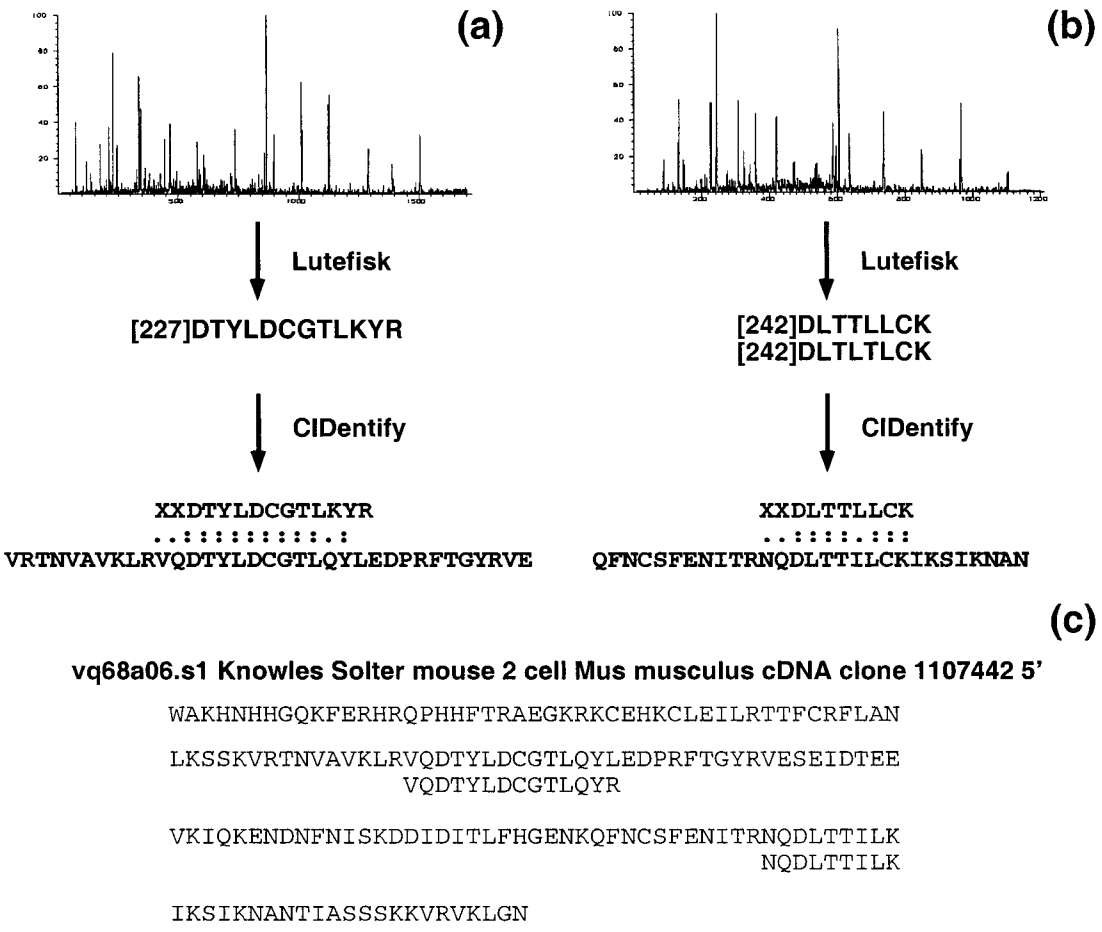
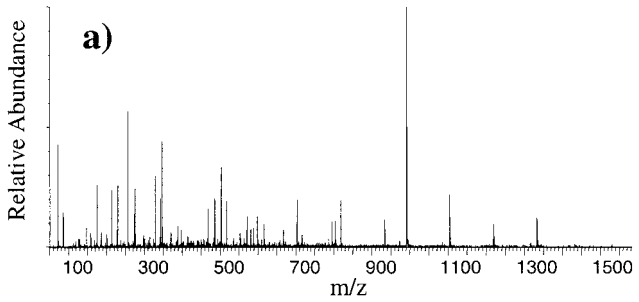


Figure 5. Interspecies match between human peptides and a mouse EST sequence. (a) Lutefisk determined a single sequence candidate that when used as input for CIDentify to search the EST nucleotide database a nonidentical mouse sequence was identified. Comparison of the deduced human sequence with the translated mouse sequence shows only a single amino acid change, where the arginine that constitutes the C-terminal tryptic cleavage site in the human sequence is replaced by leucine in the mouse sequence. (b) A second peptide was found to be an exact match to a peptide found in the same mouse EST sequence described in the first panel. (c) The translated amino acid sequence of the mouse EST sequence is shown with the two human tryptic peptide sequences aligned underneath.

tide (carbamidomethylated Cys−Asp) differed by 17 u compared to the database-derived sequence.

If the modification results in an amino acid with the same mass as a dipeptide jump, Lutefisk can correctly determine the sequence of a peptide containing a modified residue remote from the N-terminus. This was the case for a peptide obtained from a biotinylated protein purified from an avidin column (Figure 6). No database match could be made; however, using the Lutefisk results shown in Figure 6b, a homology-based database search was performed using CIDentify, and the top match was to an avidin-derived tryptic peptide—SSVNDIGDDWKATR. Hence, this potentially interesting "unknown" spectrum was quickly shown to be due to avidin column bleed. The manufacturer of the avidin beads would not confirm or deny the presence of acetylated lysine in their product; nevertheless, it was clear that this was the case. The sequence variants produced by Lutefisk (Figure 6b) are typical for Qtof data, where the candidates differ only slightly. In this case, there were two dipeptide masses of 170.10 and 170.11 u that distinguished the top two sequences. Other sequence variants determined by Lutefisk were due to tryptophan being replaced by dipeptides of masses 186.06 and 186.07 u. The N-terminal dipeptide had a mass that could only be ascribed to



b) [SS]VNDLGDDW[170.11]ATR
[SS]VNDLGDDW[170.10]ATR
[SS]VNDLGDD[186.06][170.10]ATR
[SS]VNDLGDD[186.07][170.11]ATR

Figure 6. Identification of a chemically modified peptide. (a) Qtof MS/MS spectrum of a tryptic peptide of $(M + 2H)^{2+}$ *m/z* 803.4. (b) Lutefisk sequence candidates determined from the spectrum in panel a. These results were used to search a nonredundant protein sequence database, where it was found that the best match was to an avidin peptide—SSVNDIGDDWKATR. The most likely explanation is that this peptide contains an acetylated lysine.

two serines, which is denoted as [SS]. In this example, the lysine at position 11 was replaced by a dipeptide of mass 170, which also corresponds to a single residue of acetylated lysine. Since the mass of acetylated lysine happens to be the same as the mass of two common amino acids (Ala plus Val or Leu plus Gly), Lutefisk was able to sequence this modified peptide without prior knowledge of the possibility of an acetylated lysine.

Automated de novo sequencing can also identify peptides resulting from nonconsensus protease cleavages. An example of this is peptide 9 in Table 3, which resulted from a nontryptic cleavage at the N-terminus of the peptide. Assuming that the peptide could only be derived from tryptic cleavages, a database search of this peptide revealed no matches. Although Lutefisk did not give completely correct results, in this case, a homology-based search (CIDentify) using the partially correct sequence candidates gave the correct identification (casein) as the top match.

CONCLUSIONS

We have made a number of modifications to the automated de novo sequencing program, Lutefisk. It now takes into account the extensive b-type fragmentations observed in ion trap instruments and has also been altered so that it can effectively deal with the temperature-dependent mass drift associated with the time-of-flight analyzer in the Qtof. As a result, glutamine and lysine can be differentiated, even when absolute errors of 0.2 u are observed.

In an age of extensive and expanding sequence databases, we still find automated de novo sequencing to be useful. We have found that automated de novo sequencing can be used to electronically validate database matches. In conjunction with a homology-based database search program (CIDentify), it is possible to identify homologous protein families from MS/MS data obtained from unknown proteins. Finally, automated de novo sequencing can identify posttranslational or chemical modifications, as well as peptides derived from nonconsensus protease cleavages.

Received for review October 11, 2000. Accepted March 15, 2001.

AC001196O