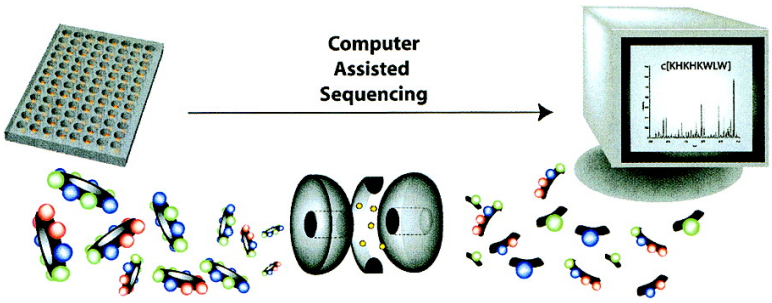
JOURNAL OF combinatorial CHEMISTRY

Article

Automated Mass Spectrometric Sequence Determination of Cyclic Peptide Library Members

James E. Redman, Keith M. Wilcoxen, and M. Reza Ghadiri

J. Comb. Chem., **2003**, 5 (1), 33-40• DOI: 10.1021/cc0200639 • Publication Date (Web): 14 December 2002



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 6 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Automated Mass Spectrometric Sequence Determination of Cyclic Peptide Library Members

James E. Redman, Keith M. Wilcoxen, and M. Reza Ghadiri*

Departments of Chemistry and Molecular Biology and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California 92037

Received July 31, 2002

Cyclic peptides have come under scrutiny as potential antimicrobial therapeutic agents. Combinatorial split-and-pool synthesis of cyclic peptides can afford single compound per well libraries for antimicrobial screening, new lead identification, and construction of quantitative structure−activity relationships (QSAR). Here, we report a new sequencing protocol for rapid identification of the members of a cyclic peptide library based on automated computer analysis of mass spectra, obviating the need for library encoding/decoding strategies. Furthermore, the software readily integrates with common spreadsheet and database packages to facilitate data visualization and archiving. The utility of the new MS-sequencing approach is demonstrated using sonic spray ionization ion trap MS and MS/MS spectrometry on a single compound per bead cyclic peptide library and validated with individually synthesized pure cyclic D,L-α-peptides.

Introduction

The proliferation of high-throughput technologies into modern drug discovery and development platforms has deepened the need for efficient combinatorial methods for the generation of diverse molecular libraries.[1] Solid phase split-and-pool synthesis[2−4] provides a rapid and manageable route to high diversity libraries but is often limited by the requirement for elaborate encoding/decoding strategies, such as chemical tagging, positional arraying, or electronic tagging.[5−7] With the advent of mild ionization techniques,[8] mass spectrometry has become an increasingly important tool for the analysis of biomolecules[9] and has been successfully applied to the deconvolution of small molecule and peptide libraries. Hit sequences from peptide libraries have classically been identified by Edman degradation, but this method is expensive and not generally applicable due to its requirement for free N termini.[10] Novel techniques have been described recently where small percentages of resin-bound peptide are "capped", providing a ladder sequence identifiable by MALDI-MS.[11,12] However, these approaches, while inexpensive and applicable to large diversity libraries, can introduce significant amounts of impurities into tested compounds. An alternative is sequence determination using post source decay[13] (PSD) or collision-induced dissociation[14] (CID) mass spectrometry, which does not require capped sequences but rather relies on the well-known fragmentation patterns of peptide bonds.[15] Linear peptides may be readily sequenced by mass spectrometry, and several automated computer programs[16−21] are available for this purpose. A number of attempts have been made to sequence cyclic peptides using mass spectrometry,[22−32] although these have not yet reached the level of automation that has been achieved for linear peptides. Homodetic cyclic peptides typically yield complex fragment ion mass spectra, arising from ring opening at multiple positions to afford mass degenerate ions, which undergo subsequent loss of residues. Only in special cases where ring opening occurs by preferential cleavage of a certain peptide bond, commonly at a proline residue due to the greater basicity of the amide nitrogen,[24,27] can the resulting ion series be interpreted in a manner similar to linear peptides. However, in many cases, further mass selection and fragmentation of linear acylium ions are necessary to fully assign a sequence,[28] requiring time-consuming interpretation of spectra. Furthermore, most automated sequencing programs are oriented toward proteomics applications and not combinatorial chemistry, so they cannot automatically take into account the constraints imposed by the chemical synthesis. We sought to develop a general automated sequence analysis platform for rapid identification of cyclic peptide library members that could be interfaced directly with existing spreadsheet or database applications for facile data mining, visualization, and archiving. Such an approach could be used to identify both active and inactive compounds—information essential for the development of accurate quantitative structure−activity relationships (QSAR). In this paper, we describe the development of such a system and validate its utility in identifying members of cyclic hexa- and octa-D,L-α-peptide libraries, a class of compounds that have shown promise as antimicrobial agents.[33]

Results and Discussion

Computer Software. The difficulty of sequencing cyclic peptide library members can be reduced if the peptide length and residue constraints imposed by a split-and-pool synthesis are taken into account. The first step of our algorithm is library enumeration from the combination of residues used at each "split" stage of the synthesis. The peptide topology

* To whom correspondence should be addressed. E-mail: ghadiri@scripps.edu.

10.1021/cc0200639 CCC: $25.00
Published on Web 12/14/2002

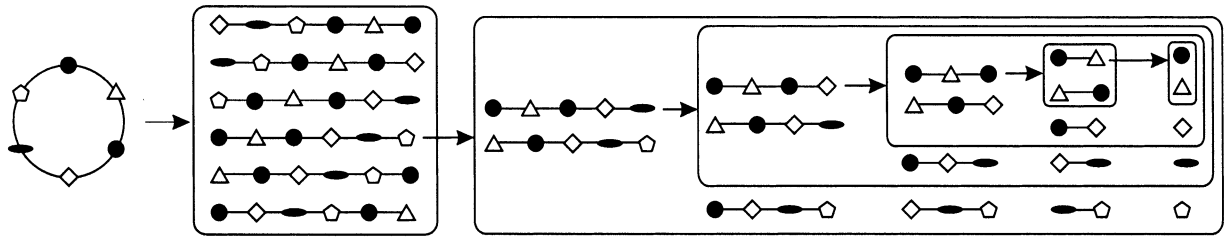


Figure 1. CID fragmentation scheme for cyclic peptides. The sequencing software considers a cyclic peptide to ring open to a mixture of linear ions, which undergo further fragmentation. Boxes and arrows indicate those fragments that originate from a common parent. Fragmentation of only a single ring-opened ion is shown for clarity.

is specified as cyclic or cyclic D,L to indicate peptides with the same or alternating chirality at the α carbon, respectively. The program generates all possible cyclic sequences,[20] stores them on a list, excluding those related by cyclic permutation symmetry. Each sequence is then associated with a score, which is calculated by comparison of an experimental mass spectrum with a spectrum predicted for that sequence. Raw experimental data, and not a peak list, are used for this comparison. The scoring is performed according to eq 1.

$$S = \frac{\sum_{j=1}^{n} w_j \sqrt{I_j}}{\sum_{j=1}^{n} w_j} \qquad (1)$$

The experimental peak intensity of peak j, I_j, is the highest intensity signal within a specified m/z range, to allow for measurement error, from a calculated peak position. Empirically, we observed that the term $\sqrt{Ij}$, as opposed to Ij, helps to reduce incorrect sequence assignment due to the occasional occurrence of intense peaks or noise spikes, which cannot be accounted for by the ions known to the software. Parameters w_j are scoring weights, which are allocated for the ion type of each peak. Normalization ensures that calculated spectra with a greater number of peaks, or comparatively more peaks with higher scoring weights, do not receive higher scores due to summation of baseline noise. In the absence of any experimental peaks, all sequences receive identical scores. After scoring of each sequence, the list is sorted, and those sequences with a score below a specified fraction of the highest scoring sequence may be discarded from the list. This process is repeated for each available experimental spectrum to ultimately yield a shortlist of candidate sequences and scores. Ordinarily, a parent ion spectrum would be considered first, followed by fragment spectra, although this is not a requirement. For the examples discussed below, this order of events was followed, and the final ranked shortlist corresponds to those sequences with a calculated parent ion mass closely matching that observed.

We make a number of simplifying assumptions in the prediction of spectra. As the cyclic peptides of interest are typically of a molecular mass <2000 Da, an extensive distribution of isotopes is not observed. We therefore consider only a single isotope peak but estimate whether it will contain 1, 2, or 3 ^{13}C atoms according to the mass of the all ^{12}C isotopomer.[19] For parent ion spectra, up to triply charged ions may be considered, with a user-specified scoring weight for each. For fragment spectra, we consider only singly charged ions and a limited set of fragmentations. Each fragmentation type is assigned a different scoring weight in eq 1. To maintain generality, no preferential sites of fragmentation are assumed and linear sequences arising from all possible fragmentations of the peptide bonds are computed, along with a map, which describes which species could be derived from each other by further fragmentation. From this sequence information, m/z is calculated for protonated fragments. These fragmentations correspond to the *b* and *y* ion types observed for linear peptides.[34] This is illustrated schematically in Figure 1. Also included are fragments derived from loss of CO and residue specific loss of NH_3 from K, Q, and R and H_2O from S, T, and E.[35] Species arising from losses of combinations of more than one NH_3 or H_2O are not considered. Multiple steps of mass selection and fragmentation can be simulated, if required, by filtering the list of m/z values followed by creation of a new list of derived fragments from the initially computed mapping. An approximation in this approach is that side chain fragmentations and loss of CO are "forgotten" after mass selection, but this has the advantage of reducing the computational burden.

An additional feature that we have termed "critical analysis" (CA) permits the calculation of comparative scores for a pair of sequences, given an experimental fragment spectrum. Scores are computed for each sequence as described above but only using calculated m/z values that are predicted to be unique to each sequence. This also reveals any potential ambiguities due to one or both sequences possessing no distinguishing peaks. This calculation is useful when applied to pairs of the highest scoring sequences on a shortlist, although it can be too time consuming to apply to all possible pairs of sequences from a diverse library.

The software is designed to be integrated into other commercially available applications, such as a spreadsheet or database, running under the Windows operating system. All of the functions described above can be accessed without the need to understand the source code or to recompile the program. As an example, we developed an Excel spreadsheet-based user interface. The processing parameters, such as spectrum filenames and scoring weights for each type of fragment, are entered directly into the spreadsheet. The user directs the program operation using a customized menu, and all results, together with a summary of the processing steps, appear in a new sheet. Various utility functions are also provided, for example, to generate a graphic representation of overlaid experimental and calculated spectra. In this way,

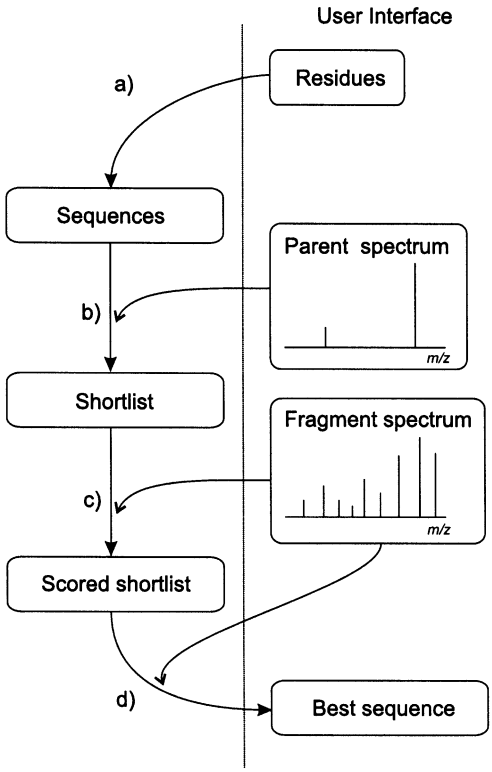


Figure 2. Flowchart of typical program operation. (a) Library member enumeration; (b) scoring of sequences against parent ion spectrum followed by elimination of sequences to form a shortlist; (c) scoring of shortlist sequences against fragment spectrum; and (d) pairwise comparison of sequences against fragment spectrum (CA).

all of the information associated with one or more sequence determinations may be contained in a single file and be readily combined with assay data. The typical flow of program operations is summarized in Figure 2.

Experimental Validation of the Mass Spectrometry (MS) Sequencing Protocol. Cyclic peptides **1**−**41** were used as the initial standard set for evaluating the performance of the automated computer-assisted MS-sequencing algorithm described above. In addition, a small test library was also designed and used to establish the utility of the MS-sequencing protocol. Peptides were synthesized and cyclized on the solid support employing Fmoc chemistry and an orthogonal resin anchoring and protecting group strategy[36,37] (Scheme 1). Standard peptides **1**−**41** were purified by crystallization and/or reversed phase high-performance liquid chromatography (RP HPLC). Single compound per bead cyclic peptide library members were synthesized on macrobead (500−550 μm) polystyrene resin. This synthetic approach allows for the rapid synthesis of a variety of cyclic peptides where cleavage of the peptide from the resin yields unprotected cyclic hexa- and octapeptides in good yields and purity required for subsequent bioassays.

The cyclic D,L-hexapeptides **1**−**11** were analyzed using an Hitachi 3DQ-LC/MS system utilizing a sonic spray ionization (SSI) interface.[38−40] SSI is an ionization method that allows for a wide range of liquid chromatography (LC) conditions, including high buffer concentration, variable compound concentrations, and thermally unstable molecules that are not readily applicable to electrospray ionization

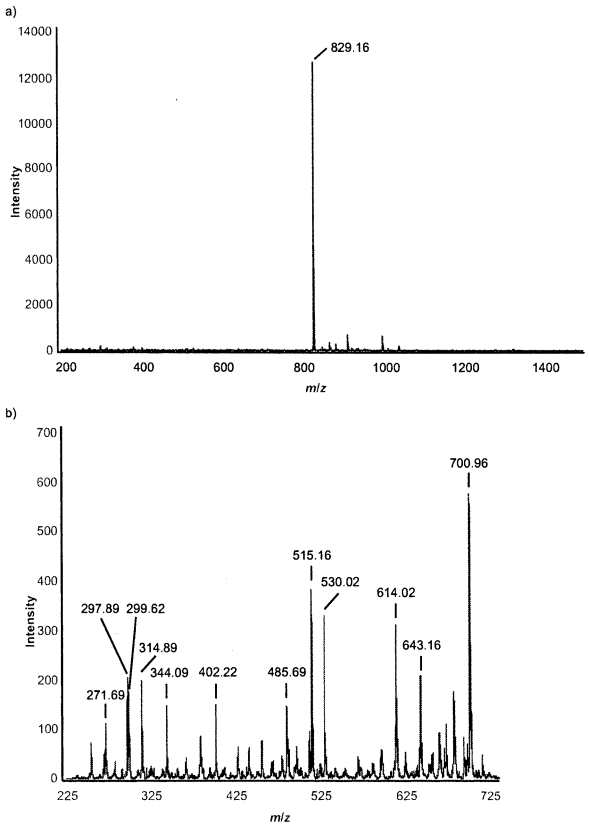
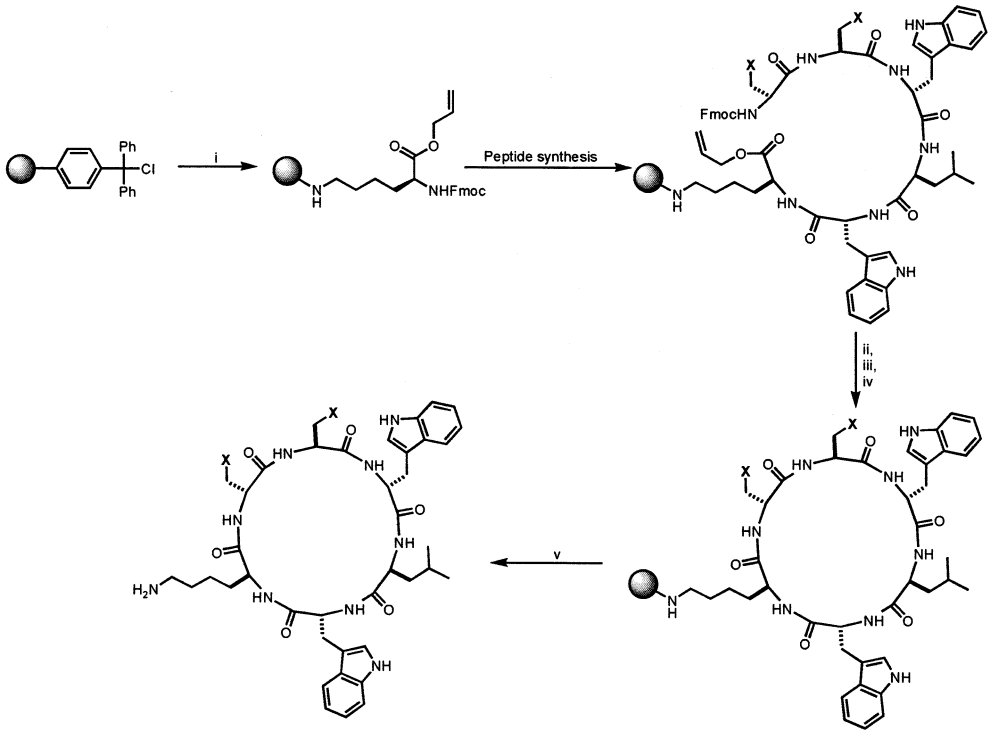


Figure 3. (a) Parent ion mass spectrum of **4**; (b) CID MS/MS spectrum of **4**.

(ESI). The difference lies in the method of charge drop formation. The ESI technique operates by applying a high voltage to the metal tip of a capillary, thereby producing fine droplets of the sample that are directed into the coupled mass spectrometer. The SSI method forms charged droplets under much milder conditions, where charged droplet formation is due to coaxial flow of nitrogen around the capillary at sonic velocity. The high-speed gas flow coupled with a small coaxial capillary diameter produces a sheer stress effect on the liquid sample to produce small charged droplets.

As compounds were individually prepared using on-resin cyclization of a protected peptide anchored through a lysine side chain, we have considered these compounds to be members of a hypothetical library in which a single residue is constrained to lysine, and the remaining five residues vary freely. However, we intentionally excluded glutamine and isoleucine from the calculations as these residues are isobaric with lysine and leucine, respectively. Therefore, the program considers 18 residues at each of the five variable positions, leading to a total of 1.8×10^6 sequences. As input, the program was supplied with a parent ion mass spectrum and CID spectrum obtained by mass selection and fragmentation of the parent and a set of processing parameters, which remained constant for every sequence. Representative parent and fragment spectra of compound **4**, c[K*W*L*W*K*S*], are shown in Figure 3, and assignments of significant fragments recognized by the software are given in Table 1. After both spectra were processed, a shortlist of scored sequences was obtained, for which CA scores were determined for all pairs of the top 30 sequences. The "best" sequence is automatically selected from the top 30 by finding that which has the fewest

Scheme 1[a]



[a] Synthesis of cyclic D,L-peptides. Reagents and conditions: (i) Fmoc-α-Lys-OAllyl (1 equiv), DIPEA (2 equiv), CH_2Cl_2, room temperature, 12 h. (ii) $Pd(PPh_3)_4$ (0.5 equiv), $CHCl_3$, 4-methylmorpholine, room temperature, 5 h. (iii) 25% Piperidine in DMF, room temperature, 2 × 10 min. (iv) PyBOP (5 equiv), HOAt (5 equiv), DIPEA (20 equiv), DMF, room temperature, 24 h. (v) TFA:H_2O:TIS (95%:2.5%:2.5%), room temperature, 12 h.

Table 1. Peak Assignments for the CID Spectrum of **4**, c[K*S*K*W*L*W*][a]

m/*z*	sequence
700.96 (701.3774)	SKWLW, WLWKS
643.16 (643.3931)	KSKWL, LWKSK
614.02 (614.35)	KWLW, WLWK
530.02 (530.3091)	KSKW, WKSK
515.16 (515.2981)	SKWL, LWKS
485.69 (486.2505)	WLW
402.22 (402.2141)	SKW, WKS
344.09 (344.2297)	KSK
314.89 (315.1821)	WK, KW
299.62 (300.1712)	LW, WL
297.89 (298.1555)	KW−NH_3, WK−NH_3
271.69 (272.1763)	WL−CO, LW−CO

[a] Calculated *m*/*z* values are given in parentheses.

"disagreements" in the CA. If more than one sequence has the same number of disagreements, then that with the highest score is selected. A sequence has a "disagreement" when both it and the sequence to which it is being compared have unique peaks in their calculated fragment spectra and the alternative sequence has a higher CA score.

The results of the analysis on hexameric sequences are summarized in Table 2. The processing parameters are such that the shortlist consists of the sequences with a molecular mass that closely matches that observed in the parent ion spectrum, ranked according to the fragment spectrum scores. For example, for compound **4**, c[K*W*L*W*K*S*] ($[M + H]^+$ 829.4725), the shortlist also contains sequences such as c[K*W*N*T*L*W*] (829.4361), c[K*W*E*W*K*A*] (829.4361), and c[K*W*R*D*R*S*] (829.4432). In all cases, the sequence predicted by the CA is identical with the known sequence, except with Q substituted with the isobaric K. As these residues also have a similar side chain fragmentation, it is difficult to distinguish between them without derivatization. The merit of the CA calculation is demonstrated by compound **6**, c[K*R*W*L*W*L*], for which the correct sequence is not the highest scoring on the shortlist but is identified as that with the fewest disagreements in the CA. The CA indicates that the retro sequence c[K*L*W*L*W*R*] must also be considered as a candidate, as this pair of sequences has no unique peaks that are known to the software.

For the diversity of the hypothetical library presented here, the full calculation takes less than 2 min per compound on a modest computer (800 MHz). In a real library, the diversity would typically be considerably less, thus speeding the calculation further. Using the Excel spreadsheet interface, the calculation and presentation of the results proceed with no user involvement.

We sought to investigate the applicability of the sequencing protocol to cyclic octapeptides to test the generality of the method for longer sequences. Stock solutions of cyclic D,L-octapeptides **12**−**41**, as prepared for biological assays,[33] were analyzed using LC/MS runs, which were representative of the analysis that would be required for a library member of unknown identity. MS/MS spectra were acquired by automatically fragmenting the most intense ion in the parent spectrum of the LC peak, in contrast to the hexamer peptides, for which MS/MS spectra were manually optimized for compounds directly infused into the spectrometer. Sequencing was performed using identical program parameters and a hypothetical library consisting of a single lysine residue and seven variable residues each with nine possible amino acids (AFHKLNRSW). In this case, the calculated library diversity is 4.0×10^6, which covers all of the 30 sequences

Table 2. Automatic Sequencing Results for Cyclic D,L-Hexapeptides[a]

compd no.	sequence	no. of sequences on shortlist	highest scoring sequence	sequence selected from CA
1	c[K*W*L*WK**E***]	1055	c[KWLWKE]	c[KWLWKE]
2	c[K*W*L*WK**F***]	714	c[KWLWKF]	c[KWLWKF]
3	c[K*W*L*WK**H***]	692	c[KWLWKH]	c[KWLWKH]
4	c[K*W*L*WK**S***]	3733	c[KWLWKS]	c[KWLWKS]
5	c[K*W*Q*W*L*W*]	201	c[KWLWKW]	c[KWLWKW]
6	c[K*R*W*L*WL]	559	c[KLWRWL]	c[KRWLWL]
7	c[K*E*Q*W*L*W*]	1055	c[KWLWKE]	c[KWLWKE]
8	c[K*F*Q*W*L*W*]	714	c[KWLWKF]	c[KWLWKF]
9	c[K*H*Q*W*L*W*]	692	c[KWLWKH]	c[KWLWKH]
10	c[K*K*Q*W*L*W*]	1221	c[KWLWKK]	c[KWLWKK]
11	c[K*Q*Q*W*L*W*]	1221	c[KWLWKK]	c[KWLWKK]

[a] CA scores were determined for the 30 highest scoring sequences on the shortlist. Residues with a D configuration at the α-carbon are italic. The stereochemical configuration of predicted sequences is not shown.

Table 3. Automatic Sequencing Results for Cyclic D,L-Octapeptides[a]

compd no.	sequence	no. of sequences on shortlist	highest scoring sequence	sequence selected from CA	rank of actual sequence
12	c[K*H*K*H*K*W*L*W*]	7145	c[KWLWKHKH]	c[KWLWKHKH]	1
13	c[K*H*Q*H*K*W*L*W*]	7145	c[KWLWKHKH]	c[KWLWKHKH]	1
14	c[K*F*Q*F*K*N*W*N*]	12 174	c[KNWNKFKF]	c[KNWNKFKF]	1
15	c[K*N*Q*N*K*F*W*F*]	12 174	c[KNKNKFWF]	c[KNKNKFWF]	1
16	c[K*A*Q*A*K*A*W*A*]	3041	c[KAWAKAKA]	c[KAWAKAKA]	1
17	c[K*A*Q*N*K*A*W*N*]	13 707	c[KNWAKNKA]	c[KWNAKNKA]	1
18	c[K*N*Q*N*K*L*W*L*]	18 720	c[KNKNKLWL]	c[KNKNKLWL]	1
19	c[K*H*K*L*A*L*W*L*]	14 298	c[KLWLALKH]	c[KLWLALKH]	1
20	c[K*S*K*L*R*L*R*L*]	17 632	c[KFLFLFKA]	c[KFAKFLFL]	2
21	c[K*S*K*L*F*L*F*L*]	21 058	c[KSKLFLFL]	c[KSKLFLFL]	1
22	c[K*H*Q*H*K*L*W*L*]	14 066	c[KLWLKHKH]	c[KLWLKHKH]	1
23	c[K*H*S*H*K*W*L*W*]	10 322	c[KWLWKHSH]	c[KWLWKHSH]	1
24	c[K*S*S*S*K*W*L*W*]	17 783	c[KWLWKSSS]	c[KWLWKSSS]	1
25	c[K*S*K*S*K*W*L*W*]	19 611	c[KWLWKSKS]	c[KWLWKSKS]	1
26	c[K*S*Q*S*K*W*L*W*]	19 611	c[KWLWKSKS]	c[KWLWKSKS]	1
27	c[*K*S*K*W*L*W*L*W*]	7204	c[KWLWLWKS]	c[KWLWLWKS]	1
28	c[K*L*W*L*W*L*W*L*]	7007	c[KLWLWLWL]	c[KLWLWLWL]	1
29	c[K*K*W*L*A*A*L*W*]	16 442	c[KKWLAALW]	c[KKWLAALW]	1
30	c[K*K*W*L*W*L*W*L*]	4772	c[KWLKWLWL]	c[KWLKWLWL]	7
31	c[K*S*K*L*W*L*W*L*]	11 898	c[KSKLWLWL]	c[KSKLWLWL]	1
32	c[*K*Q*R*W*L*W*L*W*]	2351	c[KWLWLWKR]	c[KRWKLWLW]	23
33	c[K*H*K*H*F*L*W*L*]	12 715	c[KHFKHKFH]	c[KHFLWLKH]	13
34	c[K*S*S*K*W*L*L*W*]	17 899	c[KSSKWLLW]	c[KSSKWLLW]	1
35	c[*K*K*K*W*L*W*L*W*]	3579	c[KWLWLWKK]	c[KWLWLWKK]	1
36	c[*R*H*K*K*L*W*L*W*]	6757	c[KLWRHKWL]	c[KKLWLWRH]	28
37	c[R*H*K*H*R*W*L*W*]	1325	c[KHRWLWRH]	c[KHHRWLWR]	1
38	c[R*S*K*K*L*W*L*W*]	11 089	c[KRWRKKSK]	c[KWRKRKSK]	>30
39	c[R*S*K*S*R*W*L*W*]	11 871	c[KSRWLWRS]	c[KRWLWRSS]	1
40	c[*K*Q*K*K*L*W*L*W*]	10 764	c[KWLLWKKK]	c[KKWLWKLK]	21
41	c[K*W*K*W*K*W*L*W*]	387	c[KWLWKWKW]	c[KWLWKWKW]	1

[a] CA scores were determined for the 30 highest scoring sequences on the shortlist. The rank of the actual sequence on the shortlist is indicated. Residues with a D configuration at the α-carbon are italic. The stereochemical configuration of predicted sequences is not shown.

tested, with the exception of Q, which will be interpreted as K. Despite the greater diversity of this hypothetical library, and consequentially the larger number of sequences with identical masses, the correct sequence (allowing for K/Q substitution), or its retro or enantiomeric sequence, was ranked the highest scoring for 23 compounds (Table 3). This represents an accuracy of 77%. The CA calculation selects the correct sequence in 22 cases. Interestingly, for **33** and **36**, the correct sequence is chosen even though it lies someway down the ranked list of sequences but in other cases (**17**, **37**, and **39**) an incorrect sequence is selected despite the correct sequence possessing the highest score. In only three cases (**20**, **33**, and **38**) are sequences with the wrong residue composition ranked top or selected by the CA.

The software currently lacks the ability to distinguish between a sequence and its retro sequence and between enantiomers. Although the latter problem is a fundamental limitation, these difficulties can be overcome by careful selection of residues in the library design. Likewise ambiguities between K/Q and I/L are avoidable. If only the identification of a "hit" compound is required, then several candidate sequences from the shortlist could be individually synthesized and assayed.

Sequencing Members of the Macrobead Library. A small hexamer library was constructed where two positions were varied using five amino acids in a split-and-pool fashion. The target sequence was c[K*X*X*W*L*W*] where X was a variable position containing either a Lys, His, Ser, Leu, or

Table 4. Sequences Determined for Peptides Cleaved from Single Beads of a 25 Member Library c[K*X*X*WLW*], Where X = K, H, S, W, or L

bead	sequence[a]	no. of sequences on shortlist	predicted sequence
1	c[K*L*S*W*L*W*]	71	c[K*L*S*W*L*W*]
2	c[K*S*H*W*L*W*]*	32	c[K*S*H*W*L*W*]
3	c[K*S*L*W*L*W*]*	71	c[K*S*L*W*L*W*]
4	c[K*K*S*W*L*W*]*	60	c[K*K*S*W*L*W*]
5	c[K*L*K*W*L*W*]	54	c[K*L*K*W*L*W*]
6	c[K*L*S*W*L*W*]*	71	c[K*L*S*W*L*W*]
7	c[K*L*K*W*L*W*]	62	c[K*L*K*W*L*W*]
8	c[K*H*L*W*L*W*]*	51	c[K*H*L*W*L*W*]
9	c[K*H*L*W*L*W*]*	51	c[K*H*L*W*L*W*]
10	c[K*S*L*W*L*W*]*	71	c[K*S*L*W*L*W*]
11	c[K*K*H*W*L*W*]*	12	c[K*K*H*W*L*W*]
12	c[K*K*S*W*L*W*]*	60	c[K*K*S*W*L*W*]
13	c[K*H*K*W*L*W*]*	12	c[K*H*K*W*L*W*]
14	c[K*S*S*W*L*W*]*	172	c[K*S*S*W*L*W*]
15	c[K*K*S*W*L*W*]*	60	c[K*K*S*W*L*W*]
16	c[K*S*K*W*L*W*]*	60	c[K*S*K*W*L*W*]
17	c[K*H*H*W*L*W*]*	5	c[K*H*H*W*L*W*]
18	c[K*W*W*W*L*W*]*	5	c[K*W*W*L*W*W*]
19	c[K*W*L*W*L*W*]*	15	c[K*W*L*W*L*W*]
20	c[K*H*L*W*L*W*]	51	c[K*H*L*W*L*W*]
21	c[K*K*S*W*L*W*]*	60	c[K*K*S*W*L*W*]
22	c[K*S*W*W*L*W*]*	42	c[K*S*W*W*L*W*]
23	c[K*S*S*W*L*W*]*	172	c[K*S*S*W*L*W*]
24	c[K*K*H*W*L*W*]*	12	c[K*K*H*W*L*W*]
25	c[K*W*S*W*L*W*]	42	c[K*W*S*W*L*W*]
26	c[K*S*H*W*L*W*]*	32	c[K*S*H*W*L*W*]
27	c[K*S*H*W*L*W*]*	32	c[K*S*H*W*L*W*]
28	c[K*W*W*W*L*W*]*	5	c[K*W*W*L*W*W*]
29	c[K*W*K*W*L*W*]	14	c[K*W*K*W*L*W*]
30	c[K*L*K*W*L*W*]	62	c[K*L*K*W*L*W*]

[a] An asterisk indicates that the order of the X residues could be assigned with certainty by visual inspection of the fragment spectrum or is inherently unambiguous. Automatic sequencing was carried out assuming a much larger library of 24 500 members, c[K*X*X*Z*Z*Z*], where X = A, F, G, H, K, L, M, N, R, S, T, V, W, or Y and Z = A, F, L, V, or W. The predicted sequence from the CA of the top 30 members of the shortlist is quoted.

Trp residue, and the italic residues were in a D configuration. This strategy provided a library comprising 25 individual sequences. The synthesis was carried out on 200 mg of trityl chloride macrobead resin thereby ensuring a near 100-fold redundancy of target sequences. MS and MS/MS spectra for peptides cleaved from 30 randomly selected beads were acquired in an automated hands-off manner. As the library consists of only 25 members, the two variable residues can be identified using the parent ion mass alone. In a majority of cases, the order of these two residues could also be readily determined by inspection of the fragment ion spectrum. However, to validate the software for sequencing general libraries of cationic amphiphilic cyclic peptides,[33] we considered the library to be composed of those peptides with a fixed K residue, two adjacent positions comprising neutral polar, positively charged, or hydrophobic residues (AFGHKLMNRSTVWY), and a three residue hydrophobic segment (AFLVW). This hypothetical library, of which the true library is a subset, consists of 24 500 members. Automated sequencing was carried out with the same scoring weights used for the compounds **1**−**11**, and the software predicted vs manually assigned sequences are displayed in Table 4. The number of sequences shortlisted on the basis of the observed parent ion mass are also given. An asterisk indicates those sequences that are assigned with a very high degree of confidence by visual inspection of the fragment ion spectrum. Assignment of the other sequences is based on the absence of peaks predicted to be observed for the alternative sequence in which the two variable residues are interchanged. In all cases except c[K*W*W*W*L*W*], the software selects the sequence that is believed to be correct, thus obtaining over 90% accuracy. In every case, the sequence selected by the CA calculation was identical to the highest scoring sequence.

The calculations were repeated, except using the impractically large library consisting of a single fixed lysine and five variable residues comprising all amino acids except glutamine and isoleucine. In this case, the sequences of 17 members were correctly determined, and an additional six sequences were the second or third highest scoring on the final shortlist. The most common error was the exchange of a pair of adjacent residues. Those sequences containing the fragments SWL or SLW were sometimes misassigned to permutations of the isobaric fragments EKE or EAW.

Conclusions

We have developed a system that enables rapid identification of the sequences of cyclic peptide library members from split-and-pool synthesis using sonic spray ionization CID ion trap mass spectrometry and computer analysis of spectra. There is no requirement for chemical encoding, which simplifies library preparation and avoids contamination of the library products with the encoding molecules. Our software has been designed with high-throughput screening in mind and integrates with other application programs to provide seamless transfer of data and a customizable user interface. Validation was carried out using 11 individually synthesized cyclic D,L-hexapeptides and 30 cyclic D,L-octapeptides of known sequence. Aside from ambiguities due to isobaric residues and retro sequences, the correct hexapeptide sequences were identified in all cases from a hypothetical library with a single residue constrained to lysine. Three-quarters of the cyclic octapeptides were correctly identified from a hypothetical library of over 4 million compounds. A judicious choice of residues in the library design would eliminate ambiguity, although when it exists, the software makes this apparent. The software was also tested for sequencing cyclic hexapeptides cleaved from single beads of a small split-and-pool library and demonstrated high accuracy at selecting the true sequence from a larger hypothetical library of cationic amphiphilic cyclic hexapeptides.

Experimental Section

General. Acetonitrile (HPLC grade), dichloromethane (optima grade), diethyl ether (anhydrous), dimethylformamide (DMF, sequencing grade), diisopropylethylamine (DIPEA, peptide synthesis grade), and piperidine (anhydrous) were purchased from Fisher and used without further purification. Trifluoroacetic acid (TFA, New Jersey Halocarbon), 2-(1-*H*-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU, Novabiochem), and benzotriazole 1-yl-oxy-tris-pyrrolidino-phosphonium hexaflu-

orophosphate (PyBOP, Novabiochem) were used without further purification. Tetrakistriphenylphosphine palladium-(0) was purchased from Strem Chemicals. Commercially available *N*-Fmoc amino acids for solid phase peptide synthesis and trityl chloride PS (1% DVB, substitution 0.5−1.05 mmol g^{-1}) resin were used as obtained from Novabiochem or Bachem. Trityl chloride macrobead resin was obtained from Peptides International.

Fmoc-L-Lysine(Boc)-OAllyl was made according to the protocol of Kates et al.[37] Fmoc-Lys-(Boc)-OH (5 g, 10.6 mmol) was added to allyl bromide (25 mL, 0.29 mol), followed by DIPEA (3.73 mL). This mixture was heated at 90 °C for 1 h. The reaction was allowed to cool and was concentrated by rotary evaporation, and after it was diluted with ethyl acetate, it was washed with 2 × 0.1 N HCl and 2 × saturated sodium bicarbonate at pH < 9.5, followed by brine. The organic layer was filtered through a pad of silica gel and concentrated to afford a solid. This solid was washed with ether to provide a white powder that was used directly in the next step.

Deprotection of Boc Group. An appropriate amount of Fmoc-Lys(Boc)-OAllyl for a resin loading of 0.5 mmol g^{-1} was placed in a round-bottom flask. Sufficient dichloromethane to dissolve the solid was added followed by an equivalent amount of TFA. After it was stirred for 1 h, the solution was evaporated and the residue of Fmoc-Lys-OAllyl was dried in vacuo.

Resin Loading. Trityl chloride resin was swollen in dry deacidified (Na_2CO_3) dichloromethane for 20 min. A solution of crude Fmoc-Lys-OAllyl in dichloromethane was added to the resin, immediately followed by 4 equiv of DIPEA. After the resin was shaken for 2 h, the resin was washed with dichloromethane, and then shaken with 10% MeOH: 10% DIPEA:80% dichloromethane for 10 min. After it was washed with dichloromethane and dried in vacuo, the resin loading was evaluated based on Fmoc released monitored by UV absorption at 290 nm.

Peptide Synthesis. Compounds **27**, **32**, and **33** were available from previous studies.[33] Compound **28** was prepared using standard Boc chemistry on PAM resin, employing Boc-Lys(Fmoc)-OH, cleavage with HF, cyclization in DMF solution using HATU, followed by deprotection of the Lys side chain. The remaining peptides were synthesized using standard solid phase Fmoc protocols[41] on the Fmoc-Lys-OAllyl-loaded trityl resin or for **13** and **22** using Fmoc-Glu-OAllyl-loaded Rink resin.[42] Following synthesis of the linear peptide, the resin was swollen in dry dichloromethane for 20 min. To the resin was added a degassed solution of 0.5 equiv of $Pd(PPh_3)_4$ in 90% $CHCl_3$:10% 4-methylmorpholine. After it was shaken under Ar for 5 h, the resin was washed with a solution of 1% sodium dimethylthiocarbamic acid in DMF (3 × 2 min) and 1% DIPEA (3 × 2 min) in DMF. After the final Fmoc deprotection (25% piperidine in DMF, 2 × 10 min), the resin was washed thoroughly with DMF (3 × 3 min), 10% DIPEA/DMF (3 × 3 min), and 0.8 M LiCl/DMF (3 × 3 min). The resin was treated with 5 equiv of PyBOP, 5 equiv of HOAt, and 20 equiv of DIPEA in 0.8 M LiCl/DMF for at least 12 h. After it was washed with DMF (3 × 3 min) and DCM (2 × 3 min) followed by MeOH, the peptide was cleaved from the resin and deprotected with 2.5% TIS:2.5% H_2O:95% TFA. Peptides were recovered by precipitation with ether or by evaporation of the cleavage mixture. The general properties of this class of peptides have been described elsewhere.[43]

Preparation of Split-and-Pool Library. The general synthetic protocol for library preparation was identical to that used for synthesis of individual compounds. The peptides c[K*XXWLW*], where X = K, H, L, W, or S, were prepared on trityl chloride macrobead resin (200 mg) using the split-and-pool approach.[2,3] After they were washed with Et_2O and dried in vacuo, individual beads were placed in single eppendorf tubes. To each tube was added 100 μL of a cleavage cocktail (95% TFA:2.5% H_2O:2.5% TIS) and allowed to stand for 12 h. After cleavage, the TFA mixture was evaporated and the samples were dried in vacuo to remove all volatiles (this includes TFA mixture and protecting groups, with the exception of trityl). Wells were then treated with 100 μL of 49:49:2 H_2O:CH_3CN:TFA.

Mass Spectrometry. Parent mass spectra and fragmentation spectra were obtained using a Hitachi model M-8000 3DQ-ion trap spectrometer equipped with a SSI source. Compounds **1**−**11** were dissolved in a water:acetonitrile:TFA mixture (1:1:0.01) and infused directly through the SSI source at a rate of 2 mL/hour. Parent mass peaks were selected for fragmentation using a filtered noise field (FNF). Modulation of the FNF allowed selective filtering of all ions that did not correspond to the parent ion, permitting selective fragmentation of parent peptide ion. Fragmentation spectra were acquired over an average run of 2 min, with an ion accumulation time of 500 ms, an ionization voltage of 1 kV, and collision-induced decay voltage (FNF) between 0.3 and 0.4 kV. Fragmentation spectra were obtained by averaging over the entire 2 min run. Spectra of library members from individual beads were obtained by utilizing an automated LC/MS and MS/MS run. Specifically, 50 μL of the solution containing the cleaved peptide (as outlined above) was eluted with an LC gradient of 8:2 water:acetonitrile to 3:7 water:acetonitrile over 25 min on a Vydac C18 column. Spectra were obtained using an automated MS/MS function of the Hitachi M-8000 software, where the most intense ion peak was isolated and a CID voltage was applied corresponding to the target ion mass (with other parameters similar to above). Generally, LC peak width ranged from 0.2 to 0.5 min, and spectra were obtained by averaging over this time. Because the total accumulation time was far less than that of the infusion method outlined above (2 min), the spectra were typically of an inferior signal-to-noise ratio. Spectra of compounds **12**−**41** were obtained using analogous LC/MS runs, eluting 50 μL of a solution of peptide concentration 1 mg mL^{-1} in 2.5% DMSO and 9% sucrose. An ion accumulation time of 400 ms was used for these octapeptides.

Computational. Sequencing software was developed using Borland C++ Builder and compiled to a dynamic link library (DLL). The DLL is an automation server that complies with the Microsoft Component Object Model. An Excel 2000 client was written using Visual Basic for Applications. The software was tested under Windows ME, NT4, and XP. Calculations were performed on an 800 MHz

Pentium 3 computer with 512 Mb RAM. For the calculations presented here, scoring weights, w_j, in eq 1 were 3, 2, 1, and 0 for fragments with no side chain/CO loss, loss of NH_3, loss of CO, and loss of H_2O, respectively. I_j was the highest observed intensity within ± 0.5 m/z of predicted peak j.

Acknowledgment. This work was supported by a grant from NIGMS (GM52190). J.E.R. thanks The Wellcome Trust for a postdoctoral fellowship (061454/Z/00/Z), and K.M.W. thanks the Skaggs Institute for a predoctoral fellowship.

Supporting Information Available. Mass spectra for compounds **1**−**41** and single bead library members. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Boyle, N. A.; Janda, K. D. *Curr. Opin. Chem. Biol.* **2002**, *6*, 339−346.
(2) Lam, K. S.; Salmon, S. E.; Hersh, E. M.; Hruby, V. J.; Kazmierski, W. M.; Knapp, R. J. *Nature* **1991**, *354*, 82−84.
(3) Furka, A.; Sebestyén, F.; Asgedom, M.; Dibó, G. *Int. J. Pept. Protein Res.* **1991**, *37*, 487−493.
(4) Yan, B. *Curr. Opin. Chem. Biol.* **2002**, *6*, 328−332.
(5) Czarnik, A. W. *Curr. Opin. Chem. Biol.* **1997**, *1*, 60−66.
(6) Barnes, C.; Balasubramanian, S. *Curr. Opin. Chem. Biol.* **2000**, *4*, 346−350.
(7) Affleck, R. L. *Curr. Opin. Chem. Biol.* **2001**, *5*, 257−263.
(8) Vestal, M. L. *Chem. Rev.* **2001**, *101*, 361−375.
(9) Siuzdak, G. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 11290−11297.
(10) Wang, P.; Arabaci, G.; Pei, D. *J. Comb. Chem.* **2001**, *3*, 251−254.
(11) Youngquist, R. S.; Fuentes, G. R.; Lacey, M. P.; Keough, T. *J. Am. Chem. Soc.* **1995**, *117*, 3900−3906.
(12) Hoffmann, C.; Blechschmidt, D.; Krüger, R.; Karas, M.; Griesinger, C. *J. Comb. Chem.* **2002**, *4*, 79−86.
(13) Chaurand, P.; Luetzenkirchen, F.; Spengler, B. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 91−103.
(14) Jennings, K. R. *Int. J. Mass Spectrom.* **2000**, *200*, 479−493.
(15) Polce, M. J.; Ren, D.; Wesdemiotis, C. *J. Mass Spectrom.* **2000**, *35*, 1391−1398.
(16) Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G.; Shimonishi, Y.; Takao, T. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 1867−1878.
(17) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594−2604.
(18) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(19) Ishikawa, K.; Niwa, Y. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 373−380.
(20) Sakurai, T.; Matsuo, T.; Matsuda, H.; Katakuse, I. *Biomed. Mass Spectrom.* **1984**, *11*, 396−399.
(21) Johnson, R. S.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945−957.
(22) Eckart, K.; Schwarz, H.; Tomer, K. B.; Gross, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 6765−6769.
(23) Ishikawa, K.; Niwa, Y.; Oishi, K.; Aoi, S.; Takeuchi, T.; Wakayama, S. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 395−399.
(24) Schilling, B.; Wang, W.; McMurray, J. S.; Medzihradszky, K. F. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 2174−2179.
(25) Qin, X.-Z.; Wu, Y.; Zhao, Z.; Chen, X. *J. Mass Spectrom.* **1999**, *34*, 733−740.
(26) Lin, S.; Liehr, S.; Cooperman, B. S.; Cotter, R. J. *J. Mass Spectrom.* **2001**, *36*, 658−663.
(27) Tomer, K. B.; Crow, F. W.; Gross, M. L.; Kopple, K. D. *Anal. Chem.* **1984**, *56*, 880−886.
(28) Ngoka, L. C. M.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 732−746.
(29) Ishikawa, K.; Niwa, Y.; Hatakeda, K.; Gotoh, T. *Org. Mass Spectrom.* **1988**, *23*, 290−291.
(30) Kim, S.-D.; Knoche, H. W.; Dunkle, L. D.; McCrery, D. A.; Tomer, K. B. *Tetrahedron Lett.* **1985**, *26*, 969−972.
(31) Gross, M. L.; McCrery, D.; Crow, F.; Tomer, K. B.; Pope, M. R.; Ciuffetti, L. M.; Knoche, H. W.; Daly, J. M.; Dunkle, L. D. *Tetrahedron Lett.* **1982**, *23*, 5381−5384.
(32) Aubagnac, J. L.; Devienne, F. M.; Combarieu, R. *Tetrahedron Lett.* **1983**, *24*, 2263−2266.
(33) Fernandez-Lopez, S.; Kim, H.-S.; Choi, E. C.; Delgado, M.; Granja, J. R.; Khasanov, A.; Kraehenbuehl, K.; Long, G.; Weinberger, D. A.; Wilcoxen, K. M.; Ghadiri, M. R. *Nature* **2001**, *412*, 452−455.
(34) Biemann, K. *Methods Enzymol.* **1990**, *193*, 886−887.
(35) Dookeran, N. N.; Yalcin, T.; Harrison, A. G. *J. Mass Spectrom.* **1996**, *31*, 500−508.
(36) Kunz, H.; Waldmann, H.; Unverzagt, C. *Int. J. Pept. Protein Res.* **1985**, *26*, 493−497.
(37) Kates, S. A.; Solé, N. A.; Johnson, C. R.; Hudson, D.; Barany, G.; Albericio, F. *Tetrahedron Lett.* **1993**, *34*, 1549−1552.
(38) Hirabayashi, A.; Sakairi, M.; Koizumi, H. *Anal. Chem.* **1994**, *66*, 4557−4559.
(39) Hirabayashi, A.; Sakairi, M.; Koizumi, H. *Anal. Chem.* **1995**, *67*, 2878−2882.
(40) Hirabayashi, Y.; Hirabayashi, A.; Takada, Y.; Sakairi, M.; Koizumi, H. *Anal. Chem.* **1998**, *70*, 1882−1884.
(41) Wellings, D. A.; Atherton, E. *Methods Enzymol.* **1997**, *289*, 44−67.
(42) Polaskova, M. E.; Ede, N. J.; Lambert, J. N. *Aust. J. Chem.* **1998**, *51*, 535−540.
(43) Hartgerink, J. D.; Granja, J. R.; Milligan, R. A.; Ghadiri, M. R. *J. Am. Chem. Soc.* **1996**, *118*, 43−50.

CC0200639