Published in final edited form as:
J Proteome Res. 2011 October 7; 10(10): 4505–4512. doi:10.1021/pr200323a.

Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases

Hosein Mohimani[†], **Wei-Ting Liu**[‡], **Joshua S. Mylne**[¶], **Aaron G. Poth**[¶,§], **Michelle L. Colgrave**[§], **Dat Tran**[‖,⊥,#], **Michael E. Selsted**[‖,⊥,#], **Pieter C. Dorrestein**[‡,@], and **Pavel A. Pevzner**[*,Δ]

[†]Department of Electrical and Computer Engineering, UC San Diego

[‡]Department of Chemistry and Biochemistry, UC San Diego

[¶]Institute for Molecular Bioscience, The University of Queensland, Brisbane

[§]Division of Livestock Industries, CSIRO, Brisbane

[‖]Department of Pathology and Laboratory Medicine, School of Medicine, UC Irvine

[⊥]Center for Immunology, UC Irvine

[#]Department of Pathology and Laboratory Medicine, Keck School of Medicine, USC

[@]Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego

[Δ]Department of Computer Science and Engineering, UC San Diego

Abstract

Hundreds of ribosomally synthesized cyclopeptides have been isolated from all domains of life, the vast majority having been reported in the last 15 years. Studies of cyclic peptides have highlighted their exceptional potential both as stable drug scaffolds and as biomedicines in their own right. Despite this, computational techniques for cyclopeptide identification are still in their infancy, with many such peptides remaining uncharacterized. Tandem mass spectrometry has occupied a niche role in cyclopeptide identification, taking over from traditional techniques such as nuclear magnetic resonance spectroscopy (NMR). MS/MS studies require only picogram quantities of peptide (compared to milligrams for NMR studies) and are applicable to complex samples, abolishing the requirement for time-consuming chromatographic purification. While database search tools such as Sequest and Mascot have become standard tools for the MS/MS identification of linear peptides, they are not applicable to cyclopeptides, due to the parent mass shift resulting from cyclization, and different fragmentation pattern of cyclic peptides. In this paper, we describe the development of a novel database search methodology to aid in the identification of cyclopeptides by mass spectrometry, and evaluate its utility in identifying two peptide rings from *Helianthus annuus*, a bacterial cannibalism factor from *Bacillus subtilis*, and a θ-defensin from *Rhesus macaque*.

Keywords

Mass Spectrometry; Database Searching; Cyclopeptides

[*] To whom correspondence should be addressed Phone: (858) 822-4365. Fax: (858) 534-7029, ppevzner@ucsd.edu.

Supporting Information Available: The supporting material for this article is available free of charge via the Internet at http://pubs.acs.org.

Introduction

A growing number of cyclic peptides (cyclopeptides) that are biosynthesized by a ribosomal pathway have been discovered in recent years[1] (Figure 1 and Figure S1). The cyclic nature of the backbone renders cyclopeptides impervious to the action of exopeptidases and provides protection in some cases from endoproteases. The cyclic backbone also imparts rigidity on these molecules, which may facilitate conformation-specific interactions with other proteins. A large proportion of cyclopeptides represent biologically important agents, such as antibiotics (e.g. subtilosin A from *Bacillus subtilis*,[2,3] microcin J25 from *Escherichia coli*[4] and Circulin A and B from *Bacillus circulans*[5,6]), innate immune system peptides (e.g. θ-defensins from *Macaca mulatta*[7]), bacteriocins (e.g. uberolysin from *Streptococcus uberis*[8] and carnocyclin from *Carnobacterium maltaromaticum*[9]), toxins (e.g. amatoxin and virotoxin from *Amanita* family[10,11]), protease inhibitors (e.g. SFTI-1 from *Helianthus annuus*[12]), bacterial cannibalism agents (e.g. SKF from *Bacillus subtilis*[13,14]), agents active in plant defence (e.g. Kalata B1 from *Oldenlandia affinis*,[15,16] Cyclopsychotride A from *Psychotria longipes*[17] and Circulins from *Chassalia parvifolia*[18]) and many others. It seems that the world of ribosomal cyclopeptides is much more diverse than originally anticipated, and their structural diversities are only just beginning to be appreciated.[1] The availability of genomes for many species and our incomplete knowledge of the biosynthetic pathways employed by ribosomally synthesized cyclopeptides encourages us to use genome mining approaches in combination with mass spectrometry to discover novel cyclopeptides.

Sequencing cyclopeptides, once a heroic effort, remains a challenge today. Tandem mass spectrometry (MS/MS) provides an attractive alternative to 2D nuclear magnetic resonance (NMR) spectroscopy, as it can provide access to peptide sequence information from picograms of non-purified material.[20] However, the development of algorithms for the interpretation of mass spectra of cyclopeptides is still in its infancy. Non-ribosomal cyclopeptides are not encoded by nucleotide sequence in a genome through synthesize via mRNA to peptide. Instead, they are biosynthesized by large enzyme modules (nonribosomal peptide-synthetase), where each enzyme module is responsible for incorporating one amino acid subunit. Therefore mass spectrometrists must often conduct *de novo* interpretation of mass spectra.[21,22] *De novo* peptide sequencing algorithms give promising results for short (up to 10 amino acid) cyclopeptides,[22] but often fail to correctly sequence longer peptides.

Currently, peptide sequence tag[23] (PST)-based searches of genomes are the method of choice for sequencing longer ribosomally-synthesized peptides from mass spectrometry data. For example, using imaging mass spectrometry in conjunction with a five amino acid PST (LPHPA) search, Liu *et al*.[14] identified an active metabolites from the *Bacillus subtilis* cannibalism system. This metabolite was identified as a 26 amino acid peptide named sporulation killing factor (SKF). The success of the PST approach was critically dependent on the existence of a long series of consecutive ions with standard amino acid mass differences. The sequence tag is used to search against a database comprising proteins from the organism of interest. In the reported example, the sequence tag (LPHPA) yielded a single match when searched against the *Bacillus subtilis* proteome, however, the same tag could have many more matches if searched against larger proteomes. When a human proteome is queried with the same (LPHPA) for instance, 12 putative peptide matches results. For many species, the complete genome and hence proteome are not known and it is necessary to search against closely related species or larger databases to identify novel peptides. This implies a need for database search tools that can identify cyclopeptides, analogous to Sequest[24] and Mascot[25] for linear peptides. Recently, Colgrave *et al*.[26] proposed a method for the identification of known and novel cyclotides, a class of three-disulfide knotted plant cyclopeptides of 28-37 amino acids, by searching spectra of their linear derivatives against a database of all linearized products for all cyclotides from the Cybase database.[19] However, to date, no such database search method exists for the interrogation of genomes and proteomes.

Although most ribosomal cyclopeptides are formed via a head-to-tail ligation of a single peptide, the θ-defensins are generated by concatenation and cyclization of a pair of peptides from different proteins[7] (Figure 2). In contrast with linear peptide identification tools such as Sequest and Mascot, a cyclopeptide database search tool must also address concatenation events in addition to the more commonly observed head-to-tail ligation events.

In this paper we present Cycloquest, the first database search algorithm for cyclopeptides. The search strategy is validated using sunflower trypsin inhibitor-1 (SFTI-1) and SFTI-like 1 (SFT-L1) from *Helianthus annuus*, sporulation killing factor (SKF) from *Bacillus subtilis*, and Rhesus θ-defensin (RTD-1) from *Rhesus macaque*. Our Cycloquest software for identifying cyclic peptides from their mass spectra is open source and available at http://proteomics.ucsd.edu.

Materials and Methods

Spectral datasets

Preparation of MALDI matrix (SFTI-1 and SFT-L1)—A saturated solution of α-cyano-4-hydroxycinnamic acid (CHCA; Sigma Aldrich) was prepared by dissolving the matrix in 50% acetonitrile, 0:1% trifluoroacetic acid (TFA) with 5 mM ammonium phosphate to a final concentration of 5 mg/mL. The solution was vortexed thoroughly, sonicated in a water bath for several minutes, and centrifuged at 18,000 × g for 10 minutes at room temperature. The supernatant was used in the preparation of samples for MALDI-TOF MS.

Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (SFTI-1 and SFT-L1)—Stock solutions (~1 mg/mL) of sunflower trypsin inhibitor-1 (SFTI-1) or the peptide SFT-L1 were prepared in water and 1 μL mixed directly with the matrix (1:1, v/v). Aliquots (0.6 μL) of the mixtures were spotted on a 192 well plate (Applied Biosystems) and air dried. Mass analysis was carried out in positive ion reflector mode on a 4700 Proteomics Analyzer (Applied Biosystems) using a 200 Hz frequency tripled Nd:YAG laser operating at 355 nm. Fifty spectra at each of twenty randomly selected positions were accumulated per spot between 800 and 5000 Da using an MS positive ion reflectron mode acquisition method. MS/MS spectra were acquired at seven different laser energy settings from 4000-7000 (in increments of 500) and the spectra with optimum fragmentation was used for cyclopeptide sequencing. Calibration of the instrument was carried out using the MSCal1 peptide standard (Sigma Aldrich). Data were analyzed on the accompanying 4000 series Explorer Software.

Electrospray ionization ion trap mass spectrometry (SKF and RTD-1)—SKF and RTD-1 were prepared to a concentration of 20 μg/mL in 50:50 methanol:water with 1% acetic acid and were then subjected to electrospray ionization on a Biversa Nanomate (Advion Biosystems, Ithaca, NY) nano-spray source (pressure: 0.3 psi, spray voltage: 1.4-1.8 kV). MS spectra were acquired on a 6.42 T Finnigan LTQ-FTICR MS or a Finnigan LTQ-MS (Thermo-Electron Corporation, San Jose, CA) running Tune Plus software version 1.0 and Xcalibur software version 1.4 SR1. For MS/MS experiment, the instrument was first autotuned on the m/z value of the ion to be fragmented. Then, the ions were isolated by the linear ion trap and fragmented by collision induced dissociation (CID) (isolation window: 3 m/z; collision energy: 30). Hundreds of MS/MS scans were acquired in centroid mode and averaged using QualBrowser software version 1.4 SR1 (Thermo). The Thermo-Finnigan

RAW files containing the average spectra were then converted to mzXML file format using the program ReAdW (tools.proteomecenter.org).

Sodium Borohydride treatment of SKF—Dethiolated SKF was prepared by dissolving 1 μg of SKF with 1.5 μg $NaBH_4$ and 1.5 μg $NiCl_2$ in 6.25 μL of 60% $MeOH$. This reaction was incubated at 50°C, and an additional 1.5 μg of $NaBH_4$ and $NiCl_2$ were added into the reaction 5 and 10 minutes after initiation of the reaction to ensure complete conversion of SKF into dethiolated SKF. The mixture was then centrifuged for 1 min at 14,500 rpm to remove the insoluble particles and then purified by HPLC using an Agilent Eclipse XDB-C18 column running MeCN gradients or by C18 ZipTip (Millipore) following the manufacturers protocol prior to MS analysis.

PFA treatment of RTD-1

The PFA treatment was performed using a four step process: (1) Peptide sample 0.1-10 μg (equivalent to 50-2000 pmol) was vacuum dried; (2) 19 volumes of 97% formic acid were mixed with 1 volume of hydrogen peroxide and allowed to stand on ice for 60 min; (3) 10μL of this reagent was added to the dried sample and incubated for 30 min at room temperature; and (4) the resulting solution was vacuum dried and washed three times with 50 μl of ice cold water.

Cycloquest algorithm

Similar to the MS/MS database search algorithms employed by Sequest and Mascot, our database search consist of four steps: filtering the database (e.g. by parent mass as in Sequest or Mascot, by PST as in InsPecT, etc), constructing the theoretical spectra for candidate peptides, scoring the theoretical spectra against the experimental spectra, and finally, listing the top scoring peptide spectrum matches (PSMs). While the first and the last steps of our method are very similar to Sequest and Mascot, construction of the theoretical spectra and their scoring needed to be redefined for cyclopeptides. Fortunately, we could use the scoring defined in[21,22] with slight modifications in the second and third steps of the algorithm. Another difference between Cycloquest and major database search algorithms is that Cycloquest uses a non-enzyme search strategy. The reason for this is two-fold. Many cyclic peptides are resistant to enzymatic digestion because of their compact and often disulfide-bonded nature. Additionally, digestion of cyclic peptides may result in formation of peptide fragments too small to analyze and too difficult to confidently identify. An additional step in cyclopeptide identification is to decide whether the spectrum is generated by a cyclic or a linear peptide. We address this additional complication in the Results section.

We defined a (linear) *subpeptide* of a cyclopeptide $Peptide = A_1A_2 \cdots A_k$ as a (continuous) linear substring $A_i \cdots A_j$ of the peptide (we assume $A_i \cdots A_j = A_i \cdots A_k A_1 \cdots A_j$ in the case $j < i$). There are $k(k - 1)$ subpeptides of a peptide of length k. The mass of a subpeptide is the sum of masses of all its amino acids. We define the *theoretical spectrum* of a peptide, $\Delta(Peptide)$, as the multiset of $k(k - 1)$ subpeptide masses. For example, the theoretical spectrum of a cyclopeptide AGPT = (71.037 Da, 57.021 Da, 97.052 Da, 101.047 Da) consists of 12 masses (57.021 Da, 71.037 Da, 97.052 Da, 101.047 Da, 128.058 Da, 154.073 Da, 172.084 Da, 198.099 Da, 225.110 Da, 229.105 Da, 255.120 Da, and 269.136 Da). We represented the experimental spectrum as a set of top t high intensity masses from the spectra, where t is a parameter. $CyclicScored_\delta$ $(Peptide, S)$, the number of elements (masses) shared between theoretical spectrum of $Peptide$ and S within tolerance δ was defined (Text S1).

Distinguishing cyclopeptide spectra from linear peptide spectra

One of the challenges in identification of cyclopeptides is being able to distinguish between the spectra of linear and cyclic peptides. In this section we describe a method that given a spectrum and a protein database, enables the determination of whether the spectrum was derived from a cyclic or a linear peptide.

In addition to the $CyclicScored_\delta$ (*Peptide*, *S*) defined above, given a linear experimental spectrum *S* and a peptide *Peptide*, we define the $LinearScored_\delta$ (*Peptide, S*) as the number of masses shared between *S* and the *linear theoretical spectrum* of *Peptide* within the accuracy δ, where the linear theoretical spectrum is the set of $k - 1$ b-ions and $k - 1$ y-ions of *Peptide* of length *k* (for CID spectra).

By using the cyclic and linear scores defined above, cyclopeptides can be distinguished from linear peptides based on the normalized score. Normalization is required due to different statistics of linear and cyclic scores (Figure S2). Moreover, peptides with different length have different statistics. Therefore, we normalize the score based on structure type (cyclic or linear) and peptide length. The normalized score of a match is equal to the difference of score of that match and average score of all the matches with the same length and peptides mass within 0.5 Da tolerance, over the standard deviation of all such matches.

MS/MS database search for concatenated peptides

The θ-defensin peptides are more difficult to identify than other cyclopeptides, because they are generated by concatenation and cyclization of peptides from two different protein precursors. It is computationally difficult to score the concatenation of each peptide pair over the entire macaque proteome (with 36,424 proteins totalling to 16,143,647 amino acids).

A similar problem arises for linear peptides known as the *fusion peptide identification problem*. While Ng and Pevzner[27] proposed a method for identification of the fusion peptides, their approach is not applicable to cyclopeptides. To address the quadratic growth of the number of generated concatenates, one needs a more efficient filter than the sole parent ion mass.

Many database search methods for linear peptides are *hybrid*, meaning that they attempt to use filters constructed by *de novo* searches for PSTs in order to speed up the database search by filtration using the found PSTs.[28-31] The following subsection explains our approach for making the database search of concatenated peptides computationally feasible.

While fast implementations of linear peptide database search methods are based on PSTs, we used cycloPSTs (cyclo-Peptide Sequence Tags) to speed up our database search algorithm. Given a cycloPST $CycloPST = A_1A_2 \cdots A_k$ and a parent mass *ParentMass*, we define an artificial peptide $Peptide(CycloPST) = A_1A_2 \cdots A_kA_{k+1}$, where A_{k+1} is an artificial amino acid satisfying

$$mass\,(A_{k+1}) = ParentMass - mass\,(A_1) - \cdots - mass\,(A_k)\,.$$

For example, for *cycloPST* [156.10,57.02,99.06] corresponding to RGV and *ParentMass* = 2086:24, *Peptide*(*CycloPST*) = [156:10;57:02;99:06;1774:04]. For a cycloPST *cycloPST* and an experimental spectrum *S*, we define

$$CyclicScore_\delta\,(CycloPST, S) = CyclicScore_\delta\,(Peptide\,(CycloPST)\,, S)$$

For example, for cycloPST [RGV] from θ-defensin with parent ion mass 2086.24:

$$CyclicScore_{\delta}([RGV],S)=CyclicScore_{\delta}([156.10,57.02,99.06,1774.04],S)$$

Given an experimental spectrum S and a parent mass $ParentMass$, the first step of the algorithm consists of finding high scoring CycloPSTs. However, it is computationally difficult to try all 20^k length k cycloPSTs when k gets large. The strategy used in this study is the application of a branch and bound approach, in which all length three cycloPSTs are extended in each step by one of the 20 possible amino acids, and then a fixed number of high scoring cycloPSTs are selected for the next step. For θ-defensins, we use this approach to retain 1000 high scoring length nine cycloPSTs at each iteration. In this case, the list of 1000 high scoring cycloPSTs contained the correct cycloPST [RC*IC*RRGVC*R], where C* stands for cysteic acid, and all leucines are converted to isoleucines.

Given a high scoring cycloPST $CycloPST = A_1A_2\cdots A_k$ of length k, we can generally divide it into two parts in $k - 1$ possible ways, i.e. $\{A_1 |A_2\cdots A_k\}$, $\{A_1A_2|A_3\cdots A_k\}$, $\cdots$, $\{A_1\cdots A_{k-1}|A_k\}$. For each of these divisions, we search both fragments in the genome and select the pairs of hits that can be extended to a pair of peptides with a total mass close to $ParentMass$. Assuming peptide concatenation is N-terminal to C-terminal (excluding infeasible N-terminal to N-terminal or C-terminal to C-terminal concatenations), we only accept pairs of peptides with matching directions. By concatenating each pair of peptides, we derive a set of candidate peptides which is much smaller than the original set. The final step is scoring all the candidate peptides using the cyclic score defined. Figure 3 shows the steps of algorithm.

Results

Trypsin Inhibitor and Trypsin Inhibitor-Like peptides

The cyclopeptide SFTI-1 is a potent trypsin inhibitory peptide isolated from sunflower (*Helianthus annuus*) seeds. The peptide is 14 amino acids in length, and features a single disulfide bond and a head-to-tail cyclicized backbone.[12] The cyclic and braced nature of SFTI-1 makes the peptide more resistant to degradation than linear peptides of the same size and for this reason SFTI-1 has been extensively studied in the last decade as a potentially stable peptide-based drug template.[32] In addition to potent trypsin inhibition, SFTI-1 is shown to inhibit matriptase, a serine protease overexpressed in prostate and ovarian tumors, highlighting the importance of fast-tracking cyclic peptide discovery.[33,34] Recently, Mylne *et al*.[35] reported the identification of a 12 amino acid peptide also isolated from sunflower seeds named SFT-L1 that shares some structural elements with SFTI-1 but lacks the trypsin inhibitory activity. SFTI-1 and SFT-L1 both emerge through proteolytic processing of much larger and functionally unrelated precursor proteins. SFT-L1 was identified through similarity of its precursor PawS2 to PawS1, the precursor of SFTI-1. SFT-L1 was manually sequenced by MS/MS, and its structure was obtained by NMR.[35] In this study, we determined the sequences of these cyclopeptides by searching the six frame translation of the sunflower nucleotide database using MS/MS spectra generated by MALDI-TOF/TOF mass spectrometry.

The lack of a complete sunflower genome required that we use the Expressed Sequence Tag (EST) library of seven *Helianthus* species available at the UC Davis Compositae Genome Project website, consisting of 136,935 cDNAs (totalling 96,493,071 nucleotides). Rather than covering the whole genome, ESTs only cover the RNA coding region of genome. With our interest in ribosomally synthesized cyclopeptides, searching ESTs is entirely suitable.

Both SFTI-1 and SFT-L1 contain a single disulfide bond that interferes with collision-induced dissociation during tandem mass spectrometric analysis. The disulfide bonds were removed by reduction during sample preparation[1]. The theoretical mass to charge ratio (*m/z*) of SFTI-1 and SFT-L1 in the native form are 1513.73 and 1203.48 respectively. The theoretical *m/z* of reduced SFTI-1 and SFT-L1 are 1515.74 (observed 1515.72 Da) and 1205.49 Da (observed 1205.46 Da) respectively. The TOF spectra of SFTI-1 and SFT-L1 were collected with laser energy settings of 4500, yielding optimum fragmentation in each case to allow *de novo* sequencing and database searching. We also analyzed a synthetic linear version of SFTI-1 called SFTI-1[K,S], which corresponds to the peptide SIPPICFPDGRCTK, with reduced mass of 1533.67 Da.

The first step of the database search consisted of filtering the database by parent mass. Table 1-3 show the top scoring hits for singly charged MALDI-TOF spectra of the sunflower peptides to the six frame translation of the EST library (assuming 0.5 Da mass accuracy for the parent ion mass). After scoring MS/MS fragments, normalizing scores, and sorting, both SFTI-1[K,S] and SFT-L1 are listed as the best match, while SFTI-1 is the third top match to its spectrum. In addition to the correct peptide sequence, there are some other high scoring hits from each spectrum to the database. These hits are usually computational artifacts. An additional validation step is usually required in order to distinguish the correct sequence from the shortlisted top scoring hits, e.g. by checking if the peptide is within known protein domains, in an ER signal sequence or a non-transcribed region of genomic DNA. For large proteomics datasets, false discovery rate (FDR) of the peptide sequence matches (PSMs) can be estimated to rule out false positives, similar to what occurs in the database matching of MS data to linear peptide.

Sporulation Killing Factor

When bacteria become cannibalistic, a differentiated subpopulation harvests nutrients from their genetically identical siblings to allow continued growth in nutrient-limited conditions.[13] One of the active metabolites in *Bacillus subtilis* cannibalism is sporulation killing factor (SKF), a 26 amino acid cyclopeptide that is post-translationally modified with one disulfide and one cysteine thioether bridged to the α-position of a methionine.[14] After breaking the disulfide and thioether bridges, we were able to search for, and identify SKF in the proteome database of *Bacillus subtilis*.

The theoretical mass of SKF (with disulfide and thioether bridges) is 2781.30 Da (a triply charged ion of *m/z* 928.11 measured by FT-ICR[14] corresponds to a mass of 2781.30 Da, 1.5 ppm error). By sodium borohydride reduction, all the cysteines are reduced to alanine and all the methionines are reduced to homoalanines.[37] Sodium borohydride has no effect on any other standard amino acid. The theoretical mass of sodium borohydride reduced SKF is 2551.45 Da (a triply charged ion at *m/z* 851.49 measured by FT-ICR[14] was observed, which corresponds to a mass of 2551.44 Da, 2.2 ppm).

We use the proteome database of *Bacillus subtilis* available from UniProt with 4,188 proteins, totalling 1,230,503 amino acids.

Table 4 shows the top scoring hits for the electrospray ionization ion trap-generated spectra of the sodium borohydride reduced SKF to the *Bacillus subtilis* proteome database (assuming a 0.01 Da accuracy for the parent ion mass). The correct peptide is listed as the top scoring match.

Another active metabolites in *Bacillus subtilis* cannibalism is the killing factor (SDP), a 42 amino acid linear peptide that is post-translationally modified with a disulfide bond. We

[1] After reduction, the peptides can be alkylated to prevent reoxidation of the cysteines. The results for reduced and alkylated peptides are shown in Table S2.

analyzed a triply charged native version of SDP, with triply charged parent mass ion at 1438 Da. Cycloquest identified SDP correctly, as the top hit to the *Bacillus subtilis* proteomic database (Table 5).

θ-defensin

The first cyclopeptide discovered in animals was *θ*-defensin, an antimicrobial octadecapeptide that is expressed in the leukocytes of the *Macaca mulatta*. Like the previously characterized *α*- and *β*-defensin families, *θ*-defensins possess broad spectrum antimicrobial activities against bacteria, fungi, and protect mononuclear cells from infection by HIV-1.[38]

We were able to identify the *θ*-defensin peptide using a doubly charged ion-trap (IT)-generated tandem mass spectrum. The theoretical mass of *θ*-defensin is 2079.90 Da (a doubly charged ion at *m/z* 1040.50 was observed, which corresponds to a mass of 2080.00 Da), and after performic acid treatment it increases to 2373.70 Da (a doubly charged ion at *m/z* 1188.50 was observed using ion-trap, which corresponds to a mass of 2374.00 Da), indicating the presence of three disulfide bonds.

Under PFA treatment, cysteine residues are modified to cysteic acid with a residue mass of 150.99 Da. According to Williams *et. al*.[39] only cysteine residues are affected by the treatment, and the on-target oxidation is not complicated by reactions with H, M, W or Y amino acid containing peptides.

Table 6 shows highest scoring hits to the triply charged IT spectra (assuming 0.5 Da mass accuracy for the parent ion mass).

False discovery rate of Cycloquest

In order to calculate false discovery rate, we tested the method on the previously published *Shewanella oneidensis MR-1* spectral data set containing 14.5 million spectra. The spectra were acquired on an ion trap MS instrument (LCQ, ThermoFinnigan, San Jose, CA) using ESI. The protocol for acquiring the spectra and identifications from this data set is described in Gupta *et al*.[40] 28,377 peptides were reliably identified with false discovery rate 5% using InsPecT[28] (spectrum-level false discovery rate (FDR) is 1%). We selected 21,087 tryptic peptides with a net charge of 2, obtained one representative spectra for each of these peptides (most peptides were identified from multiple spectra), and grouped these by the length of their peptide identifications to form a test data set for each length. We will refer to the length of the InsPecT identification of a spectrum as the *spectrum length*.

Our test set is a set of 1,663 spectra with spectrum length 12. We searched this dataset against the *Shewanella* database, and the corresponding decoy database. The classic reverse databases are not good candidates for decoy databases, because the theoretical spectrum of a cyclic peptide PEPTIDE, is exactly equal to the theoretical spectrum of the reverse cyclic peptide EDITPEP. Therefore, instead of using reverse sequences, the decoy is generated by shuffling the odd amino acids a^{2i+1} with the even amino acids a_{2i}, for a protein sequence $a_1, a_2, \cdots, a_n$. For example, the protein sequence PEPTIDE is shuffled to EPTPDIE. After testing the method in this dataset using a parent ion mass accuracy of 0.5 Da and fragment ion mass accuracy of 0.5 Da, out of 1,663 spectra, the method classified 1595 of them as linear target hits, 25 as cyclic targets, 26 as linear decoys, and 17 as cyclic decoys. Figure 4 shows the number of cyclic targets, linear decoys and cyclic decoys for different number of identifications. It takes about 35 minutes for Cycloquest to search 1663 Shewanella spectra against Shewanella proteome (about 1 spectrum/second) on a 3.00 GHz Core 2 Duo CPU.

While this experiment indicates a small false positive rate for Cycloquest, we are unable to estimate false negative rate due to the unavailability of suitable spectral data sets for cyclopeptides.

Discussion

While the rate of the cyclopeptide identification has increased in recent years, computational approaches for the identification of cyclopeptides are still in their infancy. As a result, papers reporting new cyclopeptides typically discuss a single family of cyclopeptides per paper. In this study we have analyzed cyclopeptides from three different kingdoms.

We propose Cycloquest as a database search method for the identification of cyclopeptides from mass spectrometric data. The general steps of Cycloquest are similar to Sequest and Mascot. However, the scoring scheme used in Cycloquest is designed specifically for cyclopeptides. We demonstrated the utility of Cycloquest through its application to the sequencing of SFTI-1, a trypsin inhibitor from *Helianthus annuus* and a related peptide. Additionally, Cycloquest sequenced SKF, a bacterial cannibalism factor from *Bacillus subtilis*, and RTD-1, the θ-defensin from Rhesus macaque. Thus, Cycloquest is capable of correctly identifying all four of these cyclopeptides, opening a possibility of sequencing of novel cyclopeptides in future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Notes and References

1. Velasquez J, van der Donk W. Genome mining for ribosomally synthesized natural products. Curr Opin Cell Biol. 2011; 15:11–21.
2. Babasaki K, Takao T, Shimonishi Y, Kurahashi K. Subtilosin A, a new antibiotic peptide produced by Bacillus subtilis 168: isolation, structural analysis, and biogenesis. J Biochem. 1985; 98:585–603. [PubMed: 3936839]
3. Kawulka K, Sprules T, McKay R, Mercier P, Diaper C, Zuber P, Vederas J. Structure of subtilosin A, an antimicrobial peptide from *Bacillus subtilis* with unusual posttranslational modifications linking cysteine sulfurs to alpha-carbons of phenylalanine and threonine. J Am Chem Soc. 2003; 125:4726–4727. [PubMed: 12696888]
4. Salomon R, Farias R. Microcin 25, a novel antimicrobial peptide produced by *Escherichia coli*. J Bacteriol. 1992; 174:7428–7435. [PubMed: 1429464]
5. Fujikawa K, Suketa Y, Hayashi K, Suzuki T. Chemical structure of circulin A. Cell Mol Life Sci. 1965; 21:307–308.
6. Hayashi K, Suketa Y, Suzuki T. Chemical structure of circulin B. Cell Mol Life Sci. 1968; 24:656–657.
7. Tang Y, Yuan J, Osapay G, Osapay K, Tran D, Miller C, Ouellette A, Selsted M. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. Science. 1999; 286:498–502. [PubMed: 10521339]
8. Wirawan R, Swanson K, Kleffmann T, Jack R, Tagg J. Uberolysin: a novel cyclic bacteriocin produced by *Streptococcus uberis*. Microbiology. 2007; 153:1619–1630. [PubMed: 17464077]
9. Martin-Visscher L, van Belkum M, Garneau-Tsodikova S, Whittal R, Zheng J, Mc-Mullen L, Vederas J. Isolation and characterization of carnocyclin a, a novel circular bacteriocin produced by *Carnobacterium maltaromaticum* UAL307. Appl Environ Microbiol. 2008; 74:4756–4763. [PubMed: 18552180]
10. Wieland T. Poisonous principles of mushrooms of the genus *Amanita* Four-carbon amines acting on the central nervous system and cell-destroying cyclic peptides are produced. Science. 1968; 159:946–952. [PubMed: 4865716]

11. Faulstich H, Buku A, Bodenmueller H, Wieland T. Virotoxins: actin-binding cyclic peptides of *Amanita virosa mushrooms*. Biochemistry. 1980; 19:3334–3343. [PubMed: 6893271]

12. Luckett S, Garcia R, Barker J, Konarev A, Shewry P, Clarke A, Brady R. High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. J Mol Biol. 1999; 290:525–533. [PubMed: 10390350]

13. González-Pastor J, Hobbs E, Losick R. Cannibalism by sporulating bacteria. Science. 2003; 301:510–513. [PubMed: 12817086]

14. Liu W, Yang Y, Xu Y, Lamsa A, Haste N, Yang J, Ng J, Gonzalez D, Ellermeier C, Straight P, Pevzner P, Pogliano J, Nizet V, Pogliano K, Dorrestein P. Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of *Bacillus subtilis*. Proc Natl Acad Sci. 2010; 107:16286–16290. [PubMed: 20805502]

15. Gran L. On the effect of a polypeptide isolated from "Kalata-Kalata" (Oldenlandia affinis DC) on the oestrogen dominated uterus. Acta Pharmacol Toxicol. 1973; 33:400–408.

16. Saether O, C DJ, Campbell I, Sletten K, Juul J, N DG. Elucidation of the Primary and Three-Dimensional Structure of the Uterotonic Polypeptide Kalata B1. J Nat Prod. 1995; 34:4147–4158.

17. Witherup K, Bogusky M, Anderson P, Ramjit H, Ransom R, Wood T, S M. Cyclopsychotride A, a Biologically Active, 31-Residue Cyclic Peptide Isolated from Psychotria longipes. J Nat Prod. 1994; 57:1619–1625. [PubMed: 7714530]

18. Gustafson K, Sowder RI, Henderson L, Parson I, Kashman Y, Cardellina JJ, McMahon J, Buckheit RJ, Pannell L, Boyd M. Circulins A and B: novel HIV-inhibitor macrocyclic peptide from tropical tree Chassalia parvifolia. J Am Chem Soc. 1994; 116:9337–9338.

19. Mulvenna J, Wang C, Craik D. CyBase: a database of cyclic protein sequence and structure. Nucleic Acids Res. 2006; 36:192–194.

20. Li J, Vederas J. Drug discovery and natural products: end of an era or an endless frontier? Science. 2009; 325:161–165. [PubMed: 19589993]

21. Ng J, Bandeira N, Liu W, Ghassemian M, Simmons T, Gerwick W, Linington R, Dorrestein P, Pevzner P. Drug discovery and natural products: end of an era or an endless frontier? Nat Methods. 2009; 6:596–599. [PubMed: 19597502]

22. Mohimani H, Liu W, Liang Y, Gaudenico S, Fenical W, Dorrestein P, Pevzner P. Multiplex *de Novo* sequencing of peptide antibiotics. J Comput Biol. 2011 in press.

23. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem. 1994; 66:4390–4399. [PubMed: 7847635]

24. Eng J, McCormack A, Yates J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994; 5:976–989.

25. Perkins D, Pappin D, Creasy D, Cottrell J. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]

26. Colgrave M, Poth A, Kaas Q, Craik D. A new era for cyclotide sequencing. Biopolymers. 2010; 94:592–601. [PubMed: 20564007]

27. Ng J, Pevzner P. Cannibalism by sporulating bacteria. J Proteome Res. 2007; 7:89–95. [PubMed: 18173219]

28. Tanner S, Shu H, Frank A, Wang L, Zandi E, Mumby M, Pevzner P, Bafna V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem. 2005; 77:4626–4639. [PubMed: 16013882]

29. Frank A, Tanner S, Bafna V, Pevzner P. Peptide sequence tags for fast database search in mass-spectrometry. J Proteome Res. 2005; 4:1287–1295. [PubMed: 16083278]

30. Frank A. A ranking-based scoring function for peptide-spectrum matches. J Proteome Res. 2009; 8:2241–2252. [PubMed: 19231891]

31. Kim S, Gupta N, Bandeira N, Pevzner P. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Mol Cell Proteomics. 2009; 8:53–69. [PubMed: 18703573]

32. Swedberg J, Nigon L, Reid J, de Veer S, Walpole C, Stephens C, Walsh T, Takayama T, Hooper J, Clements J, Buckle A, Harris J. Substrate-guided design of a potent and selective kallikrein-related peptidase inhibitor for kallikrein. Chem Biol. 2009; 16:633–646. [PubMed: 19549601]

33. Jiang S, Li P, Lee S, Lin C, Long Y, Johnson M, Dickson R, Roller P. Design and synthesis of redox stable analogues of sunflower trypsin inhibitors (SFTI-1) on solid support, potent inhibitors of matriptase. Org Lett. 2007; 9:9–12. [PubMed: 17192072]

34. Long Y, Lee S, Lin C, Enyedy I, Wang S, Li P, Dickson R, Roller P. Synthesis and evaluation of the sunflower derived trypsin inhibitor as a potent inhibitor of the type II transmembrane serine protease, matriptase. Bioorg Med Chem. 2001; 11:2515–2519.

35. Mylne J, Colgrave M, Daly N, Chanson A, Elliott A, McCallum E, Jones A, Craik D. Substrate-guided design of a potent and selective kallikrein-related peptidase inhibitor for kallikrein. Nat Chem Biol. 2011; 7:257–259. [PubMed: 21423169]

36. Cycloquest can not specify the cyclization position.

37. with mass of 85.0527 Da and composition C_4H_7ON.

38. Venkataraman N, Cole A, Ruchala P, Waring A, Lehrer R, et al. Reawakening retro-cyclins: ancestral human defensins active against HIV-1. PLoS Biol. 2009; 7:e95. [PubMed: 19402752]

39. Williams B, Russell W, Russell D. High-throughput method for on-target performic acid oxidation of MALDI-deposited samples. J Mass Spectrom. 2010; 45:157–66. [PubMed: 19937915]

40. Gupta N, Tanner S, Jaitly N, Adkins J, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith R, Pevzner P. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res. 2007; 17:1362–1377. [PubMed: 17690205]

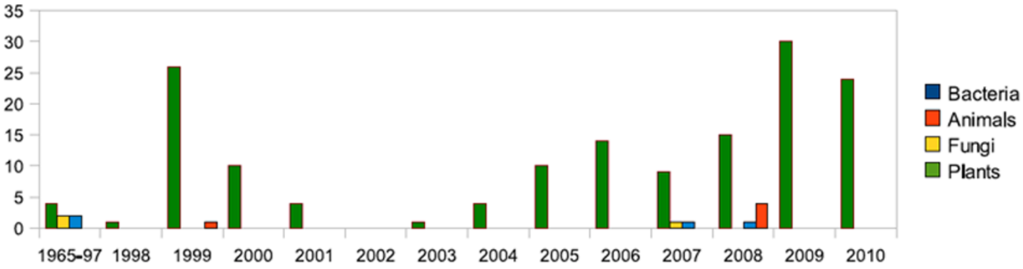

Figure 1.

Ribosomal cyclopeptides appear in all domains of life. The number of cyclopeptides sequenced in 1965-2010. The majority of known cyclopeptides have been found in plants. The data on cyclopeptides prior to 2008 have been imported from Uniprot, and the data for 2009 and 2010 have been imported from Cybase.[19] The detail of cyclopeptides found before 1997 is shown in Table S1.

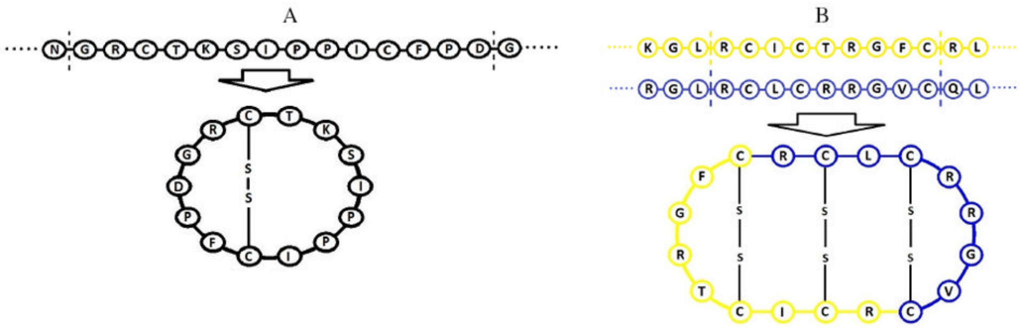
A
N G R C T K S I P P I C F P D G
C T K S I P P I C F P D G R
S
S
B
K G L R C I C T R G F C R L
R G L R C L C R R G V C Q L
C R C L C R R G V C R C I C T R G F
S
S
S
S
S
S

Figure 2.

Head-to-tail cyclization versus concatenated cyclization. (A) Head to tail cyclization of SFTI-1 from within PawS1. (B) Concatenated cyclization of *θ*-defensin. Nine amino acid segments of two different proteins (RTD1a and RTD1b) are concatenated.

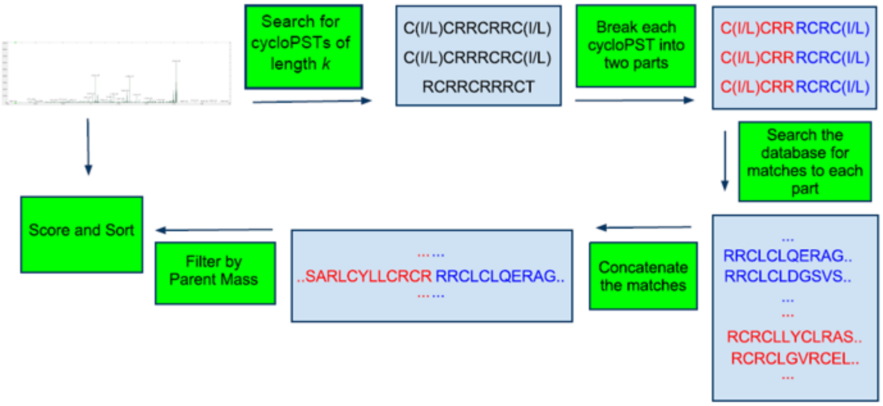

Figure 3.
Steps of Cycloquest (for concatenated cyclopeptides such as RTD-1).

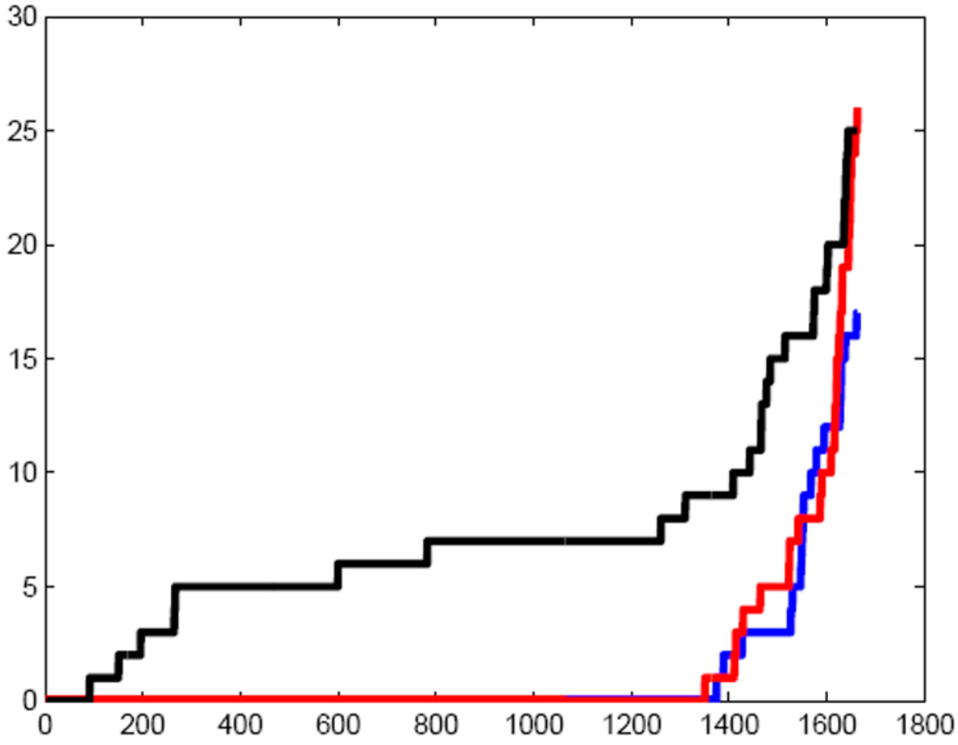

Figure 4.

The number of decoy hits plotted against the number of identification. The number of cyclic target hits is shown in black, the number of linear decoy hits is shown in red, and the number of cyclic decoy hits is shown in blue.

Table 1

Top score reconstructions for SFT-L1 from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Structure	NScore	Score	PME	cDNA clone	cDNA position	length
^{1}[a]**G**^{2}**CIEGSPVCFPD**^{1}**G**^{2}	cyclic	5.49	49	0.04	Heli-annu.bu017372-3	32	12
LRCLSVRKCQ	linear	5.19	10	0.20	Heli-annu.csa1.7627-6	283	10
CRLIFSLNHC	linear	5.19	10	0.13	Heli-annu.csa1.9966-4	168	10

[a] Superscript numerals indicate alternative cyclization positions.[36]

Table 2

Top score reconstructions for SFTI-1 from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Structure	NScore	Score	PME	cDNA clone	cDNA position	length
WRSCVGGHCNIRQ	linear	5.91	11	-0.01	Heli-para.el488692-3	180	13
QTLIHNNGINCWC	linear	5.24	10	-0.03	Heli-annu.csa1.10739-1	11	13
^{1}**G**^{2}**RCTKSIPPICFPD**^{1}**G**^{2}	cyclic	4.97	44	0.02	Heli-exil.ee660599-1	37, 38	14

Table 3

Top score reconstructions for SFTI-1[K,S] from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Structure	NScore	Score	PME	cDNA clone	cDNA position	length
SIPPICFPDGRCTK	linear	10.81	21	0.03	Heli-exil.ee660599-1	37	14
KYCLLLHRSACNL	linear	5.56	11	0.08	Heli-exil.ee644669-1	84	13
CVSSFSFSFSFWC	linear	5.56	11	-0.10	Heli-tube.csa1.1893-4	339	13

Table 4

Top score reconstruction of SKF from a triply charged mass spectrum. Correct reconstruction is shown in bold.

Peptide	Structure	NScore	Score	PME	protein	position	length
CMGCWASKSIAMTRVCALPHPAMRAI	cyclic	4.54	167	0.007	sp|p37814|atpf-bacsu	29	26
AKWLLSELNKLEKKERRKDW	cyclic	4.30	96	0.003	sp|O32215|held-bacsu	395	20
QSLKDLKGKTVGVQLGSIQEEKGK	cyclic	3.62	124	-0.008	sp|P54535|artp-bacsu	129	24

Table 5

Top score reconstruction of SDP from a triply charged mass spectrum. Correct reconstruction is shown in bold.

Peptide	Structure	NScore	Score	PME	protein	position	length
CGLYAVCVAAGYLYVVGVNAVALQTAAAVTTAVWKYVAKYSS	linear	8.17	48	0.24	sp|o34344|sdpc-bacsu	140	42
SVFFLWILNFVIGFAFPILLSSVGLSFTFFIFVALGVLA	linear	4.44	28	0.44	sp|p94493|yncc-bacsu	393	39
ELPGDLIARAQDILKELEHSGNKPEVPVQKPQVKEEPAQ	cyclic	4.05	316	0.30	sp|p49849|muts-bacsu	767	39

Table 6

Top score reconstruction for θ-defensin from a doubly charged IT spectrum. Correct reconstruction is shown in bold.

Peptide	structure	NScore	Score	PME	first protein	second protein	length
RCICTRGFCRCLCRRGVC	cyclic	15.33	160	-0.12	rhesus theta defensin-1/3	rhesus theta defensin-1/2	18
CLCRTPCNRCICTRGFCR	cyclic	13.95	149	-0.16	amiloride-sensitive cation channel 1	rhesus theta defensin-1/3	18
CRCRRCRCICTRGFCRL	cyclic	12.20	139	-0.10	hypothetical protein LOC697113	rhesus theta defensin-1/3	17