Dereplication and *de novo* sequencing of nonribosomal peptides

Julio Ng[1,7], Nuno Bandeira[2,7], Wei-Ting Liu[3], Majid Ghassemian[3], Thomas L Simmons[4], William H Gerwick[4,5], Roger Linington[6], Pieter C Dorrestein[3,5] & Pavel A Pevzner[2]

Nonribosomal peptides (NRPs) are of great pharmacological importance, but there is currently no technology for high-throughput NRP 'dereplication' and sequencing. We used multistage mass spectrometry followed by spectral alignment algorithms for sequencing of cyclic NRPs. We also developed an algorithm for comparative NRP dereplication that establishes similarities between newly isolated and previously identified similar but nonidentical NRPs, substantially reducing dereplication efforts.

The classical protein synthesis pathway (translation of template mRNA) is not the only mechanism for cells to assemble amino acids into proteins or peptides. Nonribosomal peptide synthesis is performed by nonribosomal peptide (NRP) synthetases that represent both the mRNA-free template and building machinery for the peptide biosynthesis[1]. NRP synthetases produce NRPs that are not directly inscribed in genomes and thus cannot be inferred with traditional DNA sequencing. NRPs are of great pharmacological importance as they have been optimized by evolution for chemical defense and communication. Starting from penicillin, NRPs and other natural products have an unparalleled track record in pharmacology: most anticancer and antimicrobial agents are natural products or their derivatives[2]. NRPs include antibiotics, antiviral and antitumor agents, immunosuppressors and toxins.

Most NRPs contain nonstandard amino acids, increasing the number of possible building blocks from 20 (in standard ribosomal peptides) to several hundred (**Supplementary Table 1**). Previous methods for NRP characterization are based on nuclear magnetic resonance (NMR) spectroscopy and are time-consuming and error-prone[3–5]. Therefore, there is a need for the efficient structure elucidation of NRPs. Furthermore, substantial efforts in activity screening can be saved if newly isolated compounds can be rapidly associated to a known compound by 'dereplication'[6]. Dereplication refers to the process of screening for active compounds in a mixture discarding those that have been previously studied to avoid recharacterization.

In a pioneering study[7], a cyclic algal peptide had been linearized and manually sequenced using tandem mass spectrometry (MS^2). This approach, although successful, did not result in a robust NRP sequencing technique as most NRPs evade linearization attempts. Characterization of hormothamnin A is another example of mass spectrometry–based NRP sequencing[8]. Furthermore, structural variants of antimicrobial agent tyrothricin had been characterized from a mixture of NRPs[9], using tandem mass spectrometry. In a similar experiment[10], new variations of streptocidins had also been sequenced. However, no automatic tool had been created from these studies.

We compared spectra of similar but nonidentical NRPs, enabling 'comparative dereplication' that establishes the similarity between a newly isolated and a previously identified similar (rather than identical) compounds. This is in contrast to the classical definition of dereplication, which only considers identical compounds. Because many NRPs are produced as related analogs (for example, 61 out of 90 cyanopeptides recently identified in drinking water are variants of known peptides[11]), comparative dereplication can reduce NRP characterization efforts from weeks to minutes. For example, cyanopeptide X was an unknown bioactive compound (currently known as desmethoxymajusculamide C) when our project started in 2007 but was sequenced using NMR spectroscopy in 2008. The effort invested in analyzing this NRP in 2007 would have been saved if our algorithm, NRP-dereplication, were available. Indeed, NRP-dereplication revealed that cyanopeptide X is related to majusculamide C. Another example is compound 879 that had been assumed to be new but was found to be known during the patent application. NRP-dereplication revealed that compound 879 is neoviridogrisen. NRP-dereplication derives a sequence of an unknown compound given a database of known cyclic peptides (provided a related peptide is known). In the cases when no related NRPs are known, we performed *de novo* sequencing with NRP-sequencing, a self-alignment–based algorithm, and NRP-tagging, an approach that uses frequently occurring amino acid tags for peptide reconstruction. We also reconstructed cyanopeptide X, which is to our knowledge the first report of automated *de novo* reconstruction of a cyclic peptide by mass spectrometry.

When analyzing a cyclic peptide using mass spectrometry, the MS^2 stage amounted to breaking (linearizing) the cyclic peptide into linear peptides with the same parent mass (**Fig. 1a–e**). The next stage of mass spectrometry (MS^3) breaks the different linearized versions of the cyclic peptide, resulting in the difficult problem of interpreting a MS^3 spectrum of different (but related) peptides. The theoretical MS^3 spectrum, spectrum(P) of the cyclic peptide $P = p_1 \ldots p_n$ is thus the superposition of the theoretical

[1]Bioinformatics Program, [2]Department of Computer Science and Engineering, [3]Department of Chemistry and Biochemistry, [4]Scripps Institution of Oceanography and [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA. [6]Department of Chemistry, University of California Santa Cruz, Santa Cruz, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to P.A.P. (ppevzner@ucsd.edu).

RECEIVED 25 FEBRUARY; ACCEPTED 17 JUNE; PUBLISHED ONLINE 13 JULY 2009; DOI:10.1038/NMETH.1350

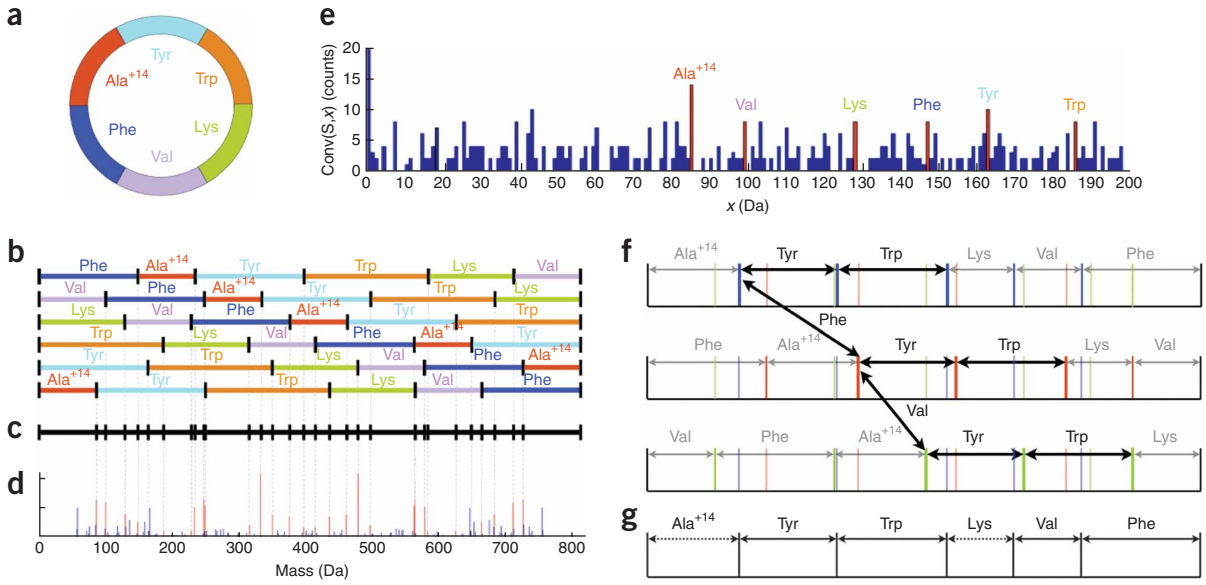

Figure 1 | Experimental and theoretical spectra of seglitide. (**a**) Cyclic diagram of seglitide. Ala^{+14} represents methylated alanine. The integer residue masses are 85, 163, 186, 128, 99 and 147 Da corresponding to cyclic A^{+14}YWKVF. (**b**) Representation of the six different theoretical linear peptides after MS^2 fragmentation of seglitide (cyclic). (**c**) Superposition of the theoretical linear fragments from **b**. (**d**) Experimental spectrum of seglitide (the peaks corresponding to fragment masses in the theoretical spectrum of segtilite in **c** are shown in red). (**e**) Autoconvolution of the spectrum in insert **d** showing prominent peaks for offsets corresponding to masses of amino acids (shown in red). The peak at 0 is truncated. (**f**) Three identical theoretical spectra of seglitide annotated as A^{+14}YWKVF (blue), FA^{+14}YWKV (red) and VFA^{+14}YWK (green) illustrating the occurrences of amino acid tags. The frequent 2-amino-acid tag Tyr-Trp was observed in three different locations in the spectrum. Additionally, the offsets between three consecutive locations of tag Tyr-Trp revealed the masses of amino acids phenylalanine and valine. (**g**) The gapped peptide constructed from **f** combines Tyr-Trp (derived from a frequent tag) with Val-Phe (derived from the inter distances between tag locations). Ala^{+14} and Lys were inferred from the flanking masses of Tyr-Trp and Val-Phe. The complete sequence A^{+14}YWKVF was recovered for seglitide, but gaps may be generated for larger compounds.

spectra spectrum(P_i) of n linear peptides $P_i = p_i \ldots p_n p_1 \ldots p_{i-1}$ for $i = 1 \ldots n$ (**Fig. 1a–e** and **Supplementary Fig. 1**).

Comparative dereplication can be formulated as the cyclic peptide dereplication problem: given an experimental spectrum S, a cyclic peptide P and a parameter k (maximum number of mutations or modifications), find a cyclic peptide P' with at most k mutations or modifications from P that maximizes the number of shared masses between S and the theoretical spectrum of P'.

We addressed the cyclic peptide dereplication problem for the most relevant case $k \leq 1$. Given the MS^3 spectrum of an unknown peptide P', and the sequence of a known peptide P that differs from P' by a single mutation at an unknown position x, NRP-dereplication derives P'. NRP-dereplication is based on the observation that most peaks shared between the experimental spectrum of P' and theoretical spectrum P correspond to subpeptides that do not contain position x (*0*-correlated subpeptides). Conversely, most peaks in the experimental spectrum P' that differ from the peaks in the theoretical spectrum of P by $\delta = \text{mass}(P') - \text{mass}(P)$ correspond to subpeptides that contain position x (δ-correlated subpeptides). The 'coverage' of a position x is defined as the number of *0*-correlated subpeptides containing that position, plus the number of δ-correlated subpeptides not containing that position. Thus, correlated subpeptides (both *0*-correlated and δ-correlated) have a potential to reveal the differing amino acid as the amino acid with the minimum coverage. For example, the drop in coverage at ornithine (**Supplementary Fig. 2**) allows one to dereplicate the experimental spectrum of tyrocidine C1 using sequence of tyrocidine C.

As the peptide P to be used for dereplication is not known in advance, every NRP spectrum needs to be compared to a database of known cyclic peptides such as Norine[12]. NRP-dereplication can localize the single mutation using the top-scoring peptide in the Norine database (**Supplementary Table 2**).

The tyrocidine family presents an ideal test for NRP-dereplication because tyrocidine A, B and C are in the Norine database, whereas tyrocidines A1, B1 and C1 are not. NRP-dereplication showed that spectra from tyrocidine A, B and C had top hits corresponding to Norine-database peptides, whereas their A1, B1 and C1 counterparts were mapped to high-scoring matches with one mutation (**Supplementary Table 2**). The correct mutated position is also localized by NRP-dereplication as the position with minimum coverage for all compounds we analyzed that had a closely related compound in the NRP database. NRP-dereplication generated only two high-scoring false hits representing very short peptides (H8495 and BQ123), but closer examination revealed that the matches were correlated to the query peptides. We conducted additional experiments that demonstrated that NRP-dereplication can localize the correct position of the mutation when $k = 1$ (**Supplementary Fig. 3**).

In the case where no related peptide is known (and thus NRP-dereplication is not applicable), we formulated the cyclic peptide sequencing problem: given an experimental spectrum S, find a cyclic peptide P maximizing the number of shared masses between S and the theoretical spectrum of P. Reconstructing the cyclic peptide P from its theoretical spectrum, spectrum(P), amounts to the cyclic version of the partial digest problem[13].

Table 1 | NRP-tagging results

Compound	Best reconstruction (masses in Da)[a]	Rank
Tyrocidine A	99, 114, 113, 147, 97, 147, 147, 114, 128, 163	3
Tyrocidine A1	99, 128, 113, 147, 97, 147, 147, 114, 128, 163	16
Tyrocidine B	99, 114, 113, 147, 97, 186, 147, 114, 128, 163	4
Tyrocidine B1	99, 128, 113, 147, 97, 186, 147, 114, 128, 163	1
Tyrocidine C	99, 114, 113, 147, 97, 186, 186, 114, 128, 163	4
Tyrocidine C1	99, 128, 113, 147, 97, 186, 186, 114, 128, 163	1
Seglitide	85, 163, 186, 128, 99, 147	1
Cyanopeptide X	57, 113, 161, 141, 71, 113, (114 + 57)[b], 127	1
BQ123	113, 186, 115, 97, 99	2
Destruxin A	113, 113, 85, 71, (98 + 97)[b]	2
H3526	97, 97, 163, 99, (97 + 1)[c], 113, (113 − 1)[c], 113	10
H8405	129, 71, 113, 113, 186	2
Microcystin LR	((83 + 71)[b] + 1)[c], (113 − 1)[c], (129 − 1)[c], (156 + 1)[c], 313, 129	27
Compound 879	113, 113, (222 − 18: 100,104)[d], (147 + 18)[c], 71, 141, 71	7
Cyclomarin A	127, 139, (286: 129,157)[d], 143, 71, (177 + 99)[b]	10
Dehydrocyclomarin A	127, 139, 268, 143, 71, 177, 99	27
Cyclomarin C	127, 139, 270, (143 + 32)[c], ((71 + 177)[b] − 32)[c], 99	>40
Dehydrocyclomarin C	Not generated	

[a]The reconstructed NRPs are represented as sequences of masses (rounded to integers) of the high-scoring peptide with a specified rank that is selected from the list of all top-scoring peptides as the most similar to the correct peptide. Actual sequenced masses are real numbers. The reconstructions given represent a complete reconstruction of the compound, or a reconstruction with composite masses and/or masses with a known offset. [b]Composite masses (2 or more amino acids). For example, 114 + 57 in cyanopeptide X means that NRP-tagging returned 171 Da as the mass of an amino acid instead of the correct masses 114 and 57 Da (corresponding to 2-hydroxy-3-methyl-pentanoic acid and glycine in cyanopeptide X). [c]Incorrect masses, expressed in terms of their offsets from correct masses. For example, 97 + 1 in H3526 means that NRP-tagging returned 98 Da whereas the correct mass is 97 Da (proline). In this case the isotopic peak (rather than a *b*-ion) was chosen as the best spectral interpretation. [d]Cases in which the algorithm split a mass, with the correct mass followed by the masses returned by the algorithm. A single mass 286 Da in cyclomarin A is split as 129 and 157 Da. A single mass 222 – 18 Da (water loss) in compound 879 is split into 100 and 104 Da.

However, it is not clear how to extend the algorithms for the partial digest problem[13,14] to a cyclic setup. Furthermore, reconstructing P from its experimental MS^3 spectrum S is a difficult problem because the contributions of different linear versions of P to the experimental spectrum are nonuniform. However, spectral convolution and spectral alignment[15] can reveal similarities between related spectra. Because an MS^3 spectrum of a cyclic peptide is a superposition of spectra of related linearized peptides, spectral autoconvolution and autoalignment reveal key features of the cyclic peptide.

Autoconvolution of a spectrum S with offset x is defined as the number of masses s in S such that $s - x$ is also a mass in S. We defined the 'cyclic' autoconvolution, conv(S,x), as the number of masses s in S such that either $(s - x)$ or $(s - x)$ + precursorMass(S) is also a mass in S. For example, high-scoring positions of the autoconvolution of seglitide revealed masses of amino acids of the NRP (**Fig. 1e**). Furthermore, the largest peak conv(S,85) = 14 corresponded to the mass of the methylated alanine (Ala^{+14}). The other five amino acids in seglitide are also represented by prominent peaks at positions 99, 128, 147, 163 and 186 with conv(S,x) ≥ 8, corresponding to their integer masses in daltons. Spectral autoconvolution (**Fig. 1e**) is a computational approach to derive residue masses of cyclic peptides.

Autoalignment of a spectrum S with offset x is defined as the set of peaks $S_x = \{s: s \in S \text{ and } (s - x) \in S\}$. Autoalignment can be viewed as a virtual spectrum with parent mass $precursorMass(S) - x$ (**Supplementary Fig. 4**). For seglitide, S_{85} ($x = 85$ maximizes conv(S,x) for seglitide) corresponds to the alignment between A^{+14}YWKVF and YWKVFA^{+14}.

Using the concepts of autoconvolution and autoalignment, we created NRP-sequencing, an algorithm to solve the cyclic peptide sequencing problem that does not require prior knowledge of the amino acid masses in the compound. NRP-sequencing first uses the MS^3 autoconvolution to derive the set of possible amino acid masses and then uses the MS^3 autoalignment using the top k possible offset masses, x, to construct a consensus spectrum S_x for each x. NRP-sequencing then generates all possible reconstructions for each S_x and reranks all generated cyclic peptides according to their matches to the MS^n spectra (for $n = 3$, 4 and 5). Details on NRP-sequencing are given in **Supplementary Note 1**. In default mode, NRP-sequencing selects the masses of the top 20 autoconvolution masses in the interval 57–200 Da and combines them with the masses of standard amino acids. NRP-sequencing could generate the correct sequence (among the set of generated reconstructions) in all cases when the resulting set of masses contained all amino acid masses in the NRP (11 out of 18 compounds). Moreover, in almost all cases the correct sequences were ranked as the top-scoring reconstruction (**Supplementary Table 3**). However, the success of NRP-sequencing is constrained by the ability to determine all amino acid masses by autoconvolution.

Because some positions are less prone to breakage than others, recovering all amino acid masses in an NRP using autoconvolution may be an unattainable goal. NRP-tagging attempts to reconstruct gapped peptides from MS^3 spectra of cyclic peptides (**Fig. 1g**). Spectra of cyclic peptides are superpositions of related (cyclically shifted) linear peptides that tend to have the same tags repeated in the spectrum. Given an MS^3 spectrum, we found all 2-amino-acid tags XY (defined by triplets of peaks s, $s + X$, $s + X + Y$ in the spectrum) and selected all frequent tags (for example, tags repeated 3 or more times). For example, if a tag XY starts at positions s, $s + A$ and $s + A + B$, then masses A and B may represent two other (adjacent) amino acids in the cyclic peptide (**Fig. 1g**). NRP-tagging first constructs a gapped peptide (for example, 85, 163, 186, 128 and 246 Da for seglitide, indicating integer masses of single or combined amino acids) and then attempts to extend it into full-length *de novo* reconstructions (for example, 85, 163, 186, 128, 99 and 147 Da, indicating integer masses of amino acids in seglitide). As gapped peptides often contain masses representing combined masses of adjacent amino acids (for example, 246 = 99 + 147 Da), NRP-tagging attempts to partition each mass in the gapped peptide into smaller masses (**Supplementary Note 2**). Similar to algorithms for sequencing linear peptides, NRP-tagging typically brings the correct peptide close to the top of the list of the high-scoring peptides (**Table 1**). This feature facilitates subsequent analysis of NRPs, for example, it allows one to correlate high-scoring reconstructions with NMR spectroscopy data. Moreover, the top-scoring peptide returned by NRP-tagging typically have minor differences as compared to the correct peptide, for example, combining masses of adjacent amino acids or choosing a mass with known offset.

Using mass spectrometry for NRP interpretation is a 'Catch-22' situation. On the one hand, there are no algorithms for interpretation of NRP spectra, thus providing little incentive for generating NRP spectra. On the other hand, shortage of NRP spectra slows down development of algorithms for NRP interpretation because spectral datasets are needed to develop such algorithms. Here we attempted to break this unfortunate cycle that will hopefully motivate the natural-product researchers to begin generating NRP spectra.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank G. Kucherov for many helpful discussions and members of the Norine team for helping with the Norine database; D. Meluzzi for the help in the data collection process; B. Moore (Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego) and A. Schultz (Scripps Institution of Oceanography, University of California, San Diego) for providing the cyclomarin compounds, and W. Fenical (Scripps Institution of Oceanography, University of California, San Diego) and K. Maloney (Scripps Institution of Oceanography, University of California, San Diego) for providing compound 879. This project was supported by US National Institutes of Health grants 1-P41-RR024851-01, GM086283 and cA10u851, and by the PhRMA foundation.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Sieber, S.A. & Marahiel, M.A. *Chem. Rev.* **105**, 715–738 (2005).
2. Newman, D.J. & Cragg, G.M. *J. Nat. Prod.* **70**, 461–477 (2007).
3. Hamada, T., Matsunaga, S., Yano, G. & Fusetani, N. *J. Am. Chem. Soc.* **127**, 110–118 (2005).
4. Ireland, C.M., Durso, A.R., Newman, R.A. & Hacker, M.P. *J. Org. Chem.* **47**, 360–361 (1982).
5. Li, J., Burgett, A., Esser, L., Amezcua, C. & Harran, P. *Angew. Chem. Int. Edn Engl.* **40**, 4770–4773 (2001).
6. Lang, G. *et al. J. Nat. Prod.* **71**, 1595–1599 (2008).
7. Krishnamurthy, T. *et al. Proc. Natl. Acad. Sci. USA* **86**, 770–774 (1989).
8. Gerwick, W.H., Jiang, Z.D., Agarwal, S.K. & Farmer, B.T. *Tetrahedron* **48**, 2313–2324 (1992).
9. Barber, M. *et al. Int. J. Mass Spectrom. Ion Process.* **122**, 143–151 (1992).
10. Hitzeroth, G., Vater, J., Franke, P., Gebhardt, K. & Fiedler, H.P. *Rapid Commun. Mass Spectrom.* **19**, 2935–2942 (2005).
11. Welker, M., Marsálek, B., Sejnohová, L. & von Döhren, H. *Peptides* **27**, 2090–2103 (2006).
12. Caboche, S. *et al. Nucleic Acids Res.* **36**, D326–D331 (2008).
13. Skiena, S.S. & Sundaram, G. *Bull. Math. Biol.* **56**, 275–294 (1994).
14. Rosenblatt, J. & Seymour, P.D. *SIAM Journal on Algebraic and Discrete Methods* **3**, 343–350 (1982).
15. Pevzner, P.A., Dancik, V. & Tang, C. *J. Comput. Biol.* **7**, 777–787 (2000).

ONLINE METHODS

Data acquisition and preprocessing. Seglitide, tyrocidines, BQ-123, destruxin A and microcystin LR were purchased from Sigma-Aldrich. H-3526 and H-8405 were purchased from Bachem. Cyanopeptide X, cyclomarins and compound 879 were provided by Gerwick's, Moore's and Fenical's laboratories at University of California, San Diego, respectively.

Time-of-flight (TOF) mass spectrometry data was acquired for tyrocidine A, A1, B, B1, C, C1; cyclomarin A, C; dehydrocyclomarin A, C; BQ123; microcystin LR; compound 879; and H8405. Ion-trap mass spectrometry data were acquired for seglitide, cyanopeptide X, destruxin A and H3526.

For the ion-trap data acquisition, each compound was prepared to 1 μM solution using 50:50 MeOH:water with 1% AcOH as solvent, and underwent nanoelectrospray ionization on a Biversa Nanomate (pressure: 0.3 p.s.i., spray voltage: 1.4–1.8 kV). Ion trap spectra were acquired on a Finnigan LTQ-MS (Thermo-Electron Corporation) running Tune Plus software version 1.0. For the MS^n data collection, spectrum ion trees were collected in both automatic mode and manual mode. In automatic mode, the [M+H]+ of each compound was set as the parent ion. MS^n data were collected with the following parameters: maximum breadth, 20; maximum MS^n depth, 3. At $n = 2$, isolation width, 4; normalized energy, 50. At $n = 3$, isolation width, 4; normalized energy 30. For manually collected data, the [M+H]+ ion of each compound was isolated with an isolation width of 3 mass to charge (m/z) units and fragmented with normalized collision energy of 30. Top 20 intense ions within the spectra were isolated with an isolation width of 3 m/z units and fragmented with normalized collision energy of 30. The Thermo-Finnigan files (in RAW format) were then converted to an mzXML file format using the ReAdW (http://tools.proteomecenter.org/).

For the TOF data collection, the cyclic peptides were prepared in a 50% methanol, 0.5% acetic acid at 1 pmol μl^{-1}. The samples were then infused into an ABI QSTAR XL QTOF using nanospray source I for ionization at 0.5 μl min^{-1}. The instrument was then set up in automatic acquisition mode to collect one MS scan to detect the calibrants (CsCl (Sigma) and cPDI inhibitor (Bachem)) and one product ion scan for the parent mass of the peptide in the experiment. Each scan time was 30 s and the method length was 2 min. The acquisition was set to enhance for the scanned ranges. The collision energy for each compound was determined using direct infusion in tune mode to find out the optimal collision energy required to produce ideal fragmentation for MS^2. The collected spectra were calibrated using the first mass spectrometry scan and the calibration was applied to the entire file.

All spectra were preprocessed before the sequencing algorithms were applied. The initial filtering steps were to ensure that the low-intensity peaks are removed. The standard procedure of keeping the top 5 peaks within a window of 50 Da was applied to all compounds.

Access to data and algorithms. All software tools and spectral annotations are available at http://bix.ucsd.edu/nrp/index.html.

doi:10.1038/nmeth.1350