Supplementary Materials For Dereplication and De Novo Sequencing of Nonribosomal Peptides

June 11, 2009

1 MS-CPA Annotation Tool

A user friendly software MS-CPA (Cyclic Peptide Annotator) was developed to annotate spectra of cyclic peptides (available at http://bioinf.ucsd.edu/~jung/ms-cpa). Given a spectrum and a candidate peptide, the tool generates a spectral annotation (see Figure 1 for an example).

2 Estimating the reliability of NRP-Dereplication in database searches

While Figure 2 demonstrates that drops in coverage reveal the differing amino acid, we need to ensure that “random” pairs of peptides do not exhibit similar drops (otherwise dereplication will fail when comparing a spectrum against a database of known NRPs). The number of correlated subpeptides for random peptide pairs is much smaller than for related peptide pairs. In each dereplication experiment, we compared the tyrocidine C experimental spectrum to an (incorrect) peptide that differed from the correct peptide by a fixed number of amino acids. While there are 32 correlated subpeptides between the experimental tyrocidine C1 spectrum and the tyrocidine C peptide (differing by a single aa), the number of these subpeptides quickly decreases as the peptides diverge (Figure 3a). Our simulations revealed that NRP-Dereplication is correct in over 90% of cases in case of a single aa difference (Figure 3). As expected, the number of correlated subpeptides drops as the number of differing amino acids increases. (see the Average Worst Rank plot in Figure 3b). Finally, Figure 4 shows the score distribution of all the matches of the experimental spectrum of compound 879 to the entries in NORINE.

3 NRP-Dereplication results

We used all cyclic and some partially cyclic peptides in NORINE in our database. The search results of our test compounds are shown in Table 1.

4 Auto-alignment

Figure 5 illustrates the notion of auto-alignment for a given offset ($S_{85}$) given the theoretical spectrum of seglitide ($A^{+14}YWKVF$). $S_{85}$ auto-alignment reveals both prefix and suffix ladders (rather than a single ladder as in the spectral alignment of linear peptides [1]). For the sake of illustration, we use a theoretical spectrum, but the concept also works for experimental spectra with noisy and missing peaks.

5 Details of NRP-Sequencing

Below we use seglitide as an example to illustrate the details of NRP-Sequencing. Similarly to the spectral alignment of spectra from different peptides [2], one would expect auto-alignment of seglitide to mostly reflect either prefix or suffix ion fragments from the linearized peptides $A^{+14}YWKVF$ and $YWKVFA^{+14}$ (with the number of noisy peaks greatly reduced). The separation of prefix and suffix ladders by spectral alignment
| Amino Acid | Mass  | Offset |
|------------|-------|--------|
| a          | 85.053| -18.011|
| Y          | 163.063| -17.027|
| W          | 186.079| 1.003 |
| K          | 128.095|        |
| V          | 99.068 |        |
| F          | 147.068|        |

Figure 1: Annotation of the experimental spectrum of seglitide. The x-axis represents m/z values, while the y-axis represents relative intensity. The number to the right of the y-axis (1.03e+6) represents the absolute intensity at 100%. The Offset column lists the neutral losses considered when annotating the spectrum.
Figure 2: Dereplication results. a) NRP-Dereplication output for experimental spectrum of tyrocidine C1 (VKLFPWWNQY) given peptide sequence of tyrocidine C (VOLFPWNQY). Concentric red-gray circles represent 0-correlated subpeptide (with peptide shown red and its complement shown gray) and δ-correlated subpeptides (with peptide shown gray and its complement shown red). Given this coloring convention, the amino acid coverage (number of red arcs covering an amino acid) represents supporting evidence that an amino acids did not change from the known to the unknown compound. The thick black circle separates 0-correlated subpeptides (shown inside) from δ-correlated subpeptides (shown outside). The outer counts represent the coverage for a given amino acid by red arcs and reveals the differing amino acid (O) as the amino acid with minimum coverage (2.5 vs. 7 for the next runner-up). The counts are normalized by the number of subpeptides per peak. For example, if a peak has two alternative subpeptide annotations, it will contribute \( \frac{1}{2} \) to the coverage. The width of the arcs are proportional to this weighting factor. The number in the center of the graphs is the total number of correlated subpeptides. b) Alternative representation of a) as a histogram that reveals the changed amino acid O. c) Additional dereplication results for the tyrocidine family.
Figure 3: Dereplication results for the experimental spectrum of tyrocidine C (VOLFPWWNQY) compared against peptides with varying number of differing amino acids. All experiments were ran 1000 times randomly choosing a different offset(s) on a random amino acid(s). Each of the differing masses have a randomly chosen offset in the range of 15 to 43 Daltons (the sign of offset is also chosen randomly).

a) The number of correlated subpeptides as a function of the number of differing amino acids.

b) The average worst rank of the known differing amino acid(s). The worst rank is the highest rank for all the differing amino acids when sorted in the increasing order of their coverage. For example, if amino acids A and B were modified, and after the dereplication experiment, their coverage ranks were 1 and 3, the worst rank would be 3. In the case of \( k \) differing amino acids, a rank of \( k \) means that the dereplication experiment was successful.

c) For case of a single amino acid difference, NRP-Dereplication algorithm is correct in over 90% of cases. We note that for the cases in which the differing amino acid has rank 2, the position with the lowest coverage is usually a neighboring position.
| Compound | Top Match(es) | Dereplicated Compound | Score |
|----------|--------------|-----------------------|-------|
| Destruxin A | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:1(3)-OH[2] [+14] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:1(3)-OH[2] [+14] | 0.45 |
| HydroxyDestruxin B | Pro, Ile, NMe-Val, NMe-Ala, bAla, iC5:0-OH[2.3]-[18] | Pro, Ile, NMe-Val, NMe-Ala, bAla, iC5:0-OH[2.3]-[18] | 0.45 |
| Destruxin D | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[20] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[20] | 0.45 |
| Destruxin E diol | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | 0.45 |
| Destruxin A | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | Pro, Ile, NMe-Val, NMe-Ala, bAla, HInv[2] | 0.45 |
| Destruxin C | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | 0.45 |
| Destruxin F | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | 0.45 |
| Destruxin B | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[4] | 0.45 |
| Destruxin E chlorohydrin | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[38] | Pro, Ile, NMe-Val, NMe-Ala, bAla, C4:0-OH[2.3.4]-[38] | 0.45 |

Table 1: NRP-Dereplication results. The Score is defined as the product of the fraction of explained intensity and the fraction of explained fragment masses of a dereplicated peptide. Dereplicated matches have monomers (shown in red) where the candidate mutation is placed with the integer mass of the offset enclosed in square brackets (Dereplicated Compounds column). See Table 4 for the complete list of monomers. Compounds that are in the database (tyrocidine A, B, C, H3526, microcystin LR and compound 879) or have a closely related compound (tyrocidines A1, B1, C1, cyanopeptide X, destruxin A) have higher scores than compounds that are not in the database (seglitide, cyclomarin A, C and dehydrocyclomarin A, C). Dereplicated compounds have the mass difference of the experimental spectrum and the mass of the peptide enclosed in square brackets next to their name (Top Matches column). The compounds are sorted by score and the double horizontal line separates compounds in the database (or have a close match) from the compounds that are not in the database (lower part of the table). Compounds H8405 and BQ123 (representing the shortest are nevertheless correlated with the correct peptide sequences. For H8405, the correct masses are [113, 71, 129, 186, 113], while the database match is [184, 186, 129, 113]. For BQ123, the correct masses are [113, 186, 115, 97, 99], while the database match is [71, 228, 71, 97, 143]. For seglitide and the family of cyclomarins, no high-scoring matches were returned by NRP-Dereplication because their sequences are not in NORINE yet. However, if we introduce any cyclomarin in NORINE, we readily dereplicate all other cyclomarins.
is important since it enables accurate de novo peptide sequencing [2, 3]. However, the interpretation of auto-alignments (of cyclic peptides) is more complex than interpretation of spectral alignments of different (linear) peptides. While auto-alignment reduces the noise, it does not separate prefix and suffix ladders, i.e., auto-alignment contains both prefix and suffix ladders. This is caused by the fact that the MS$^3$ spectrum contains peaks from both $A^{+14}YWKVF$ and $YWKVF^{+14}$. Thus, the prefix masses from $A^{+14}YWKVF$ match the prefix masses of $YWKVF^{+14}$ and, moreover, the suffix masses from $YWKVF^{+14}$ match the suffix masses of $A^{+14}YWKVF$ (with the same offset 85 for both cases).

- **Auto-convolution.** Since for cyclic peptides the set of aa-masses is not known in advance we extend the set of standard 20 aa masses using the top 20 peaks (with masses between 57 and 200 Da) in the MS$^3$ auto-convolution as the approximation for the masses of amino acids. While we implicitly assume that prominent peaks in auto-convolution correspond to masses of single amino acids, it is straightforward to extend NRP-Sequencing to the case when some of the top peaks corresponds to the mass of 2 or more amino acids. In practice, offsets from double/triple masses do not lead to the deterioration of NRP-Sequencing as long as the non-standard masses of all single amino acids are present among the top peaks (like in the case of sitgliptide).

- **Auto-alignment.** After auto-aligning the MS$^3$ spectrum with offset $x$, we construct the consensus spectrum $S_x$ containing only matching masses in the overlapping portion of the auto-alignment. The resulting consensus spectrum is scored with the summed intensities of the corresponding matched masses. For experimental spectrum of sitgliptide, the auto-alignment spectrum $S_{85}$ contains all prefix and suffix (b/y) ions for the peptide YWKVF ($x = 85$ corressponds to the most prominent peak in auto-convolution $Conv(S, x)$).

- **De novo peptide sequencing.** We solve the de novo peptide sequencing problem for the auto-alignment spectrum using the anti-symmetric path algorithm [4]. NRP-Sequencing generates all de novo peptide reconstructions of $S_x$ (for each of the top $t$ auto-convolution masses $x$, where $t$ is a parameter) with scores above $p \cdot Score(P)$, where $p$ is a parameter and $P$ is the highest scoring de novo reconstruction of $S_x$. We observed that $t = 2$ works well in most cases.

- **Re-ranking candidate peptides using MS$^n$ spectra.** NRP-Sequencing further scores each candidate peptide by matching all MS$^n$ spectra against it and re-ranking candidate peptides according to their matches to the MS$^n$ spectra. Peaks in de novo reconstructions were scored against MS$^n$ spectra using a likelihood scoring scheme as described in [5]. De novo sequences derived from TOF MS$^3$ spectra were also cyclized and scored against the MS$^3$ spectrum; MS$^3$/MS$^n$ match scores and matched peak intensities were combined using linear discriminant analysis.

Results of NRP-Sequencing are in Table 2.
Figure 5: Auto-alignment $S_{85}$ of the theoretical spectrum of seglitide (85 Da represents the most prominent peak in auto-convolution). All peaks with mass $s$ that have a related peak with mass $s + 85$ are enclosed in red circles. a) Representation of the theoretical spectrum for seglitide. Each horizontal line represents a different linear version of seglitide (prefix ladder). The linearized version $A^{+14}YWKVF$ is shown twice. The diagonal lines represent the suffix ladders. If we stay on a diagonal, and walk from left to right, we read out the mass sequence of seglitide in reverse. The vertical axis is drawn at half the scale of the horizontal axis to save space. b) Shows the cyclic theoretical spectrum of seglitide by compressing all breaks of a) into a single line. There are two set of peaks that are 85 Da apart. First, those peaks in the first and second horizontal lines (aligned prefixes) and second, those peaks in the red and cyan diagonals (aligned suffixes). These sets are highlighted with the blue and red arcs in the cyclic theoretical spectrum, respectively. c) auto-alignment of b) with offset of 85 Da. All matching peak pairs (identified by arcs) will be reflected in the consensus spectrum. The auto-alignment spectrum contains the prefix and suffix ladder of the overlapping linearized seglitide sequences. Note that in reality, prefix and suffix ladders are not separated in the auto-alignment spectrum.
| Compound         | Best reconstruction | Rank |
|------------------|---------------------|------|
| Tyrocidine A     | [163+99], 114, [113+147], [147+147], 147, [114+128] | 1    |
| Tyrocidine A1    | [163+99], 128, [113+147], [147+147], 147, [114+128] | 1    |
| Tyrocidine B     | [163+99], 114, [113+147], 97, [186+147], 114, 128 | 14   |
| Tyrocidine B1    | 99, 128, [113+147], [97+186], 147, [114+128] | 1    |
| Tyrocidine C     | 113, 147, 97, 186, 186, 114, [128+163], [99+114] | 125  |
| Tyrocidine C1    | [163+99], [128+113], 147, [97+186], 186, [114+128] | 1    |
| Selligentide     | 85, [163+186], 128, 99, 147 | 1    |
| Cyanopeptide X   | 57, 113, 161, 141, 71, [113+114+57], 127 | 1    |
| BQ123            | 113, 186, 115, [97+99] | 1    |
| H3526            | 97, [97+163], 99, [97+113], 113, 113 | 2    |
| H8405            | 129, 71, 113, 113, 186 | 1    |

Table 2: NRP-Sequencing results. The reconstructed NRPs are represented as sequences of masses. For the sake of brevity, masses are rounded to integers. Composite masses (2 or more aa) are enclosed in square brackets. For example, [163+99] in tyrocidine A means that NRP-Sequencing returned 262 (composite mass of 163 and 99 (Tyr and Val)). Best reconstruction is the highest scoring completely correct (i.e. no incorrect b-ions) de novo sequence returned by NRP-Sequencing.

| Glufib          | % Cuts | % Int | Tyrocidine C1 | % Cuts | % Int |
|-----------------|--------|-------|--------------|--------|-------|
| 65.95, 72.12, 646.54, 111.05, 342.09, 332.22 | 33     | 20    | 128.06, 146.04, 504.21, 283.12, 186.05, 114.03 | 50     | 58    |
| 156.10, 115.00, 513.55, 542.20, 160.06, 83.05 | 30     | 19    | 128.06, 163.05, 487.20, 283.12, 186.05, 114.04 | 53     | 57    |
| 64.94, 74.08, 643.78, 113.02, 341.13, 333.05 | 33     | 18    | 146.01, 99.07, 405.16, 283.13, 186.06, 242.12 | 53     | 54    |
| 97.01, 132.05, 285.42, 114.04, 656.26, 285.18 | 37     | 16    | 128.06, 163.06, 99.02, 388.19, 283.12, 300.09 | 57     | 53    |
| 57.02, 129.03, 757.67, 440.22, 114.98, 71.08 | 30     | 16    | 128.03, 163.06, 487.23, 300.10, 169.05, 114.05 | 50     | 52    |

Table 3: NRP-Tagging results for linear peptide Glufib and tyrocidine C1. % cuts is the percentage of observed fragment ions in the experimental spectrum out of all theoretical peaks in the cyclic peptide. % int is the percentage intensity explained by the annotated peaks in the spectrum (including possible neutral losses).

6 Applying NRP-Tagging to linear peptides

Below we address a possible concern that NRP-Tagging may erroneously use a spectrum of a linear peptide to dereplicate a cyclic peptide. In this section, we show that if NRP-Tagging were to be run on a mass spectrum of a linear peptide, the resulting gapped peptides would have a much lower score than those of a spectrum of a cyclic peptide. Table 3 shows the top 5 tags for a spectrum of linear peptide Glu-1-Fibrinopeptide (glufib), a standard linear peptide used for instrument calibration, of sequence EGVNDNEEGFFSAR and the top 5 tags for tyrocidine C1. Both spectra were acquired using the same experimental settings. These results indicate that NRP-Tagging may distinguish spectra from linear and cyclic peptides.
| Mass   | Name               | Name               | Mass   | Name               | Name               | Mass   | Name               | Name               | Mass   |
|--------|--------------------|--------------------|--------|--------------------|--------------------|--------|--------------------|--------------------|--------|
| 111.05 | Lys                | Glu                | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.05 | D-Leu              | dLeu               | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 111.07 | Glu                | Glu                | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
| 112.10 | D-Asp              | M-Asp              | 156.09 | D-Glu              | M-Glu              | 156.09 | D-Glu              | M-Glu              |        |
Table 4: List of masses of amino acids (monomers) in NORINE sorted by mass
(504 amino acids with 288 unique elemental compositions)

| Mass (Da) | Amino Acid(s) | Length | Mass (Da) | Amino Acid(s) |
|----------|---------------|--------|----------|---------------|
| 113.08   | D-t-Leu       | 1      | 113.08   | D-t-Leu       |
| 113.08   | Leu           | 1      | 113.08   | D-t-Leu       |
| 113.08   | D-allo        | 1      | 113.08   | Leu           |
| 113.08   | D-NMe-Val     | 1      | 113.08   | D-allo        |
| 113.08   | D-NMe-Nva     | 1      | 113.08   | D-NMe-Val     |
| 113.13   | NSPD          | 1      | 113.13   | D-NMe-Nva     |
| 114.03   | Pda           | 1      | 114.03   | NSPD          |
| 114.03   | Hap           | 1      | 114.03   | Pda           |
| 114.04   | Aen           | 1      | 114.04   | Hap           |
| 114.04   | NMe-Dpy       | 1      | 114.04   | Aen           |
| 114.07   | D-Ala         | 1      | 114.07   | NMe-Dpy       |
| 114.07   | D-Bmp         | 1      | 114.07   | D-Ala         |
| 114.07   | Hmp           | 1      | 114.07   | D-Bmp         |
| 114.07   | 4Me-D-Hva     | 1      | 114.07   | Hmp           |
| 114.07   | C6:0-OH(3)    | 1      | 114.07   | 4Me-D-Hva     |
| 114.08   | Orn           | 1      | 114.08   | C6:0-OH(3)    |
| 114.08   | D-Orn         | 1      | 114.08   | Orn           |
| 115.03   | Asp           | 1      | 115.03   | D-Orn         |
| 115.03   | D-Asp         | 1      | 115.03   | Asp           |
| 115.06   | D-hOH-Val     | 1      | 115.06   | D-Asp         |
| 115.06   | hOH-Val       | 1      | 115.06   | D-hOH-Val     |
| 115.06   | NMe-Thr       | 1      | 115.06   | hOH-Val       |
| 116.05   | D5:0-OH(2.4)  | 1      | 116.05   | NMe-Thr       |
| 116.05   | d5:0-OH(2.3)  | 1      | 116.05   | D5:0-OH(2.4)  |

References

[1] Pevzner, P.A., Dancik, V., Tang, C.: Mutation-Tolerant Protein Identification by Mass Spectrometry. J Comput Biol 7 (2000) 777–787
[2] Bandeira, N., Tsur, D., Frank, A., Pevzner, P.A.: Protein Identification by Spectral Networks Analysis. Proceedings of the National Academy of Sciences 104 (2007) 6140–6145
[3] Bern, M., Goldberg, D.: De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning. Journal of Computational Biology 13 (2006) 364–378
[4] Chen, T., Kao, M.Y., Tepel, M., Rush, J., Church, G.M.: A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. J Comput Biol 8 (2001) 325–37
[5] Dancik, V., Addona, T., Clauser, K., Vath, J., Pevzner, P.: De Novo Peptide Sequencing via Tandem Mass Spectrometry. J Comput Biol 6 (1999) 327–342