Constrained De Novo Sequencing of Peptides with Application to Conotoxins

Swapnil Bhatia[1,2], Yong J. Kil[1], Beatrix Ueberheide[3], Brian Chait[3], Lemmuel L. Tayo[4,5], Lourdes J. Cruz[5], Bingwen Lu[6,7], John R. Yates III[6], and Marshall Bern[1]

[1] Palo Alto Research Center
[2] Department of Electrical and Computer Engineering, Boston University
[3] Rockefeller University
[4] Mapua Institute of Technology, The Philippines
[5] Marine Science Institute, University of the Philippines
[6] The Scripps Research Institute
[7] Pfizer Inc.

Abstract. We describe algorithms for incorporating prior sequence knowledge into the candidate generation stage of de novo peptide sequencing by tandem mass spectrometry. We focus on two types of prior knowledge: homology to known sequences encoded by a regular expression or position-specific score matrix, and amino acid content encoded by a multiset of required residues. We show an application to de novo sequencing of cone snail toxins, which are molecules of special interest as pharmaceutical leads and as probes to study ion channels. Cone snail toxins usually contain 2, 4, 6, or 8 cysteine residues, and the number of residues can be determined by a relatively simple mass spectrometry experiment. We show here that the prior knowledge of the number of cysteines in a precursor ion is highly advantageous for de novo sequencing.

1 Introduction

There are two basic approaches to peptide sequencing by tandem mass spectrometry (MS/MS): *database search* [16], which identifies the sequence by finding the closest match in a protein database, and *de novo sequencing* [5], which attempts to compute the sequence from the spectrum alone. Database search is the dominant method in shotgun proteomics because it can make identifications from lower quality spectra with less complete fragmentation. *De novo* sequencing finds use in special applications for which protein databases are difficult to obtain. These applications include unsequenced organisms [31], biotech products such as monoclonal antibodies [3,28], phosphopeptide epitopes [13], endogenous antibodies [29], and peptide toxins [2,27,37].

In many de novo sequencing applications, partial knowledge of the sequence is relatively easy to obtain. For example, antibodies contain long conserved segments and 10- to 13-residue hypervariable segments (complementarity determining regions), and a peptide from a digest may overlap both types of regions. A fairly simple database-search program can recognize MS/MS spectra of peptides with N- or C-terminus in a known conserved segment, and these spectra can then be de novo sequenced to determine the variable segment. As another example, nerve toxins from arthropods and

V. Bafna and S.C. Sahinalp (Eds.): RECOMB 2011, LNBI 6577, pp. 16–30, 2011.

mollusks contain highly conserved cysteine scaffolds with the number and positions of the cysteines well-conserved but the other residues variable. The numbers of cysteines in various precursors can be determined by a relatively simple mass spectrometry experiment: derivatize cysteines and measure the mass shifts. In the absence of such an experiment, the researcher can simply try each guess at the number of cysteines. These two examples are by no means exhaustive. Partial knowledge may also be obtained from previous experiments or computations, sequenceable overlapping peptides, "split" isotope envelopes from certain post-translational modifications, residue-specific derivatizations, amino acid analysis, Edman sequencing, manual inspection, comparative genomics, and so forth.

In this paper we explore the possibility of using partial knowledge to guide de novo sequencing. Related previous work has used close homology to a database protein to help assemble de novo peptide sequences into a protein sequence [3] and also to correct sequencing errors [25]. We apply partial knowledge to the candidate generation stage rather than the later stages (scoring, protein assembly, and correction of mistakes) for several reasons. First, we aim to use much weaker partial knowledge, for example, the number of cysteines rather than close homology (say 90% identity) to a known sequence. Second, our partial knowledge is often exact rather than probabilistic. Third, it is logically cleaner to maintain the scorer as a function of only the candidate sequence and the mass spectrum, independent of any protein database or biological knowledge. Fourth, we find it convenient to use a single scorer (ByOnic) for all peptide identification tasks, so that we can freely combine database search and de novo sequencing results, even within a single run of the program. Almost all MS/MS data sets contain numerous spectra identifiable by database-search, from keratin and trypsin if nothing else, and leaving these spectra to be identified *de novo* reduces the number of true identifications and falsely increases the number of "interesting" de novo sequences.

The rest of the paper is organized into the following sections: problem formulation, algorithms, validation of the approach on known conotoxins, and announcement of novel conotoxins. At this point, we believe we have completely sequenced about 15 novel mature conotoxins from two species (*Conus stercusmuscarum* and *Conus textile*), but in this bioinformatics paper we report only two new toxins, one from each species, while we wait for peptide synthesis to validate our sequences. Currently only about 130 mature conotoxins (meaning exact termini and modifications) are known after 40 years of study [23], so 15 novel conotoxins represents a substantial contribution to the field. (We have also observed about 35 mature conotoxins that match database sequences in *C. textile*, slightly exceeding the original analysis of the same data sets [36,37].) Most studies add only one or two de novo sequences at a time. For example, Nair et al. [27] and Ueberheide et al. [37] each manually sequenced one novel toxin.

2 Problem Formulation

In a tandem mass spectrometer, charged peptides break into a variety of charged and neutral fragments. The mass spectrometer measures the mass over charge (m/z) of these fragments and outputs a tandem mass spectrum, a histogram of ion counts (intensities) over an m/z range from zero to the total mass of the peptide. Given a mass spectrum,

the goal of *de novo* peptide sequencing is to generate a sequence of possibly modified amino acid residues whose fragmentation would best explain the given spectrum.

Formally, a *spectrum* $\mathcal{S}$ is a triple (S, M, c) where S is a set of pairs of positive real numbers $\{(m_1, s_1), \ldots, (m_n, s_n)\}$, M is a positive real number, and c is an integer. M denotes the total mass of the peptide whose fragmentation produced S and is the sum of the masses of the amino acids in its sequence. The peptide charge is c, which is typically in the range +1 to +4 for the spectra we consider. Each pair (m_i, s_i) in S denotes a peak in the spectrum at m/z of m_i of intensity s_i. Let $\mathcal{A}$ be a set of symbols representing amino acid residues and modifications. We define a *peptide* p as a nonempty string over the alphabet $\mathcal{A}$.

We assume that we have access to a *peptide scoring function* h which, given a peptide p, spectrum $\mathcal{S}$, and a set of allowable modifications, returns the probability that $\mathcal{S}$ is produced by p. Let A be a set of distinct positive numbers representing the fixed masses of the symbols in $\mathcal{A}$. The problem of de novo *candidate generation* is this: Given an integer k (say $k = 100,000$), tandem mass spectrum $\mathcal{S}$, a set of symbols $\mathcal{A}$ and their masses A, find a set C of k candidate peptides p over the alphabet $\mathcal{A}$ such that $\max_{p \in C} h(\mathcal{S}, \mathcal{A}, A, p)$ is maximized.

The parameter k above sets a limit on the number of candidate sequences we can afford to score. We cannot afford to score all possible sequences, because the number of possible peptides of a given mass M is exponential in the length of the peptide. Prior work [2,12,30] has shown the advantage of considering sets of spectra, but in this paper we generally focus attention on the de novo sequencing of single spectra.

In accord with almost all de novo sequencing programs, such as Lutefisk [35],PEAKS [26], EigenMS [7], NovoHMM [17] and PepNovo [19], we have factored the problem into candidate generation and scoring phases. Candidate generation typically uses a dynamic programming best-path algorithm [10,11,26], to compute thousands of possible sequences. The scoring phase then scores each of these candidates, using more detailed global information such as proton mobility, fragmentation propensities, and mass measurement recalibration [7], that does not conform to the separability requirement (the "principle of optimality") of dynamic programming. Here we describe how to incorporate partial knowledge into the candidate-generation phase. For scoring, we use the scorer in ByOnic [6], which is primarily a database-search program.

De novo sequencing is well known to be a difficult problem, due to incomplete fragmentation, noise peaks, mixture spectra, and the large numbers of peptides and fragments within error tolerance for any given mass. The best de novo sequencing programs rarely give a completely correct answer on a peptide of mass 2000 Da. High-accuracy instruments [20] and CID/ETD pairs [12,30] help, yet conotoxins remain especially challenging targets due to prevalent modifications and high proline content, which tends to suppress fragmentation.

3 Constraints and Algorithms for Constrained De Novo Search

Sequence constraints restrict the search from the space of all possible peptides of the given precursor mass to a proper subset of the space, in which all peptides satisfy certain *a priori* criteria. For example, we might assume that the peptide contains 4 cysteines, as do all α-conotoxins.

To demonstrate the feasibility and utility of such a constrained search approach to de novo sequencing, and to explore its role in a de novo sequencing protocol, we implemented two types of constraints in our peptide candidate generator: a multiset constraint and a regular expression constraint. We describe these constraints and our algorithm to generate candidates satisfying them below. We also implemented a simple search algorithm, similar to SALSA [24], for searching for spectra that satisfy mass and regular expression constraints. We describe this below.

3.1 Multiset Constraint

Let $\mathcal{A}$ be the set of amino acid symbols (including modifications). A *multiset constraint* is a vector $c : \mathcal{A} \rightarrow \mathbb{N}$ describing a subset of $\mathcal{A}^\star$—the set of all strings over $\mathcal{A}$. We denote this subset by $S(c)$. Thus, the vector

$$c(\mathsf{G}) = 1;\ c(\mathsf{V}) = 2;\ c(\mathsf{C}) = 4;\ \text{ and } c(x) = 0,\ \forall x \in \mathcal{A} \setminus \{\mathsf{G}, \mathsf{V}, \mathsf{C}\};$$

is an example of a multiset constraint. A multiset constraint defines $S(c)$ in the following way: if $c(x) = n$, then x must appear at least n times in every string in $S(c)$. All x such that $c(x) = 0$ impose no constraints on $S(c)$. Thus, in the above example,

$S(c) = \{w : w \in \mathcal{A}^\star$ and w contains at least one G, at least two V, and at least four $\mathsf{C}\}$.

For example, VGCCQCPARCKCCV satisfies the constraint in the above example, but CCPARCCVR does not.

3.2 An Algorithm for Generating Multiset-Constrained Candidates

Let $\mathcal{A}$ be the set of amino acid symbols (including modifications). Let $\mathcal{S} = (T, M)$ be a given (deisotoped and decharged) spectrum, let c be a multiset constraint, and let N be a positive integer. The objective of the de novo candidate generation algorithm is to output a set of N peptides, all satisfying the multiset constraint c, containing a peptide that best explains the spectrum $\mathcal{S}$. Our algorithm proceeds in two stages. In the first stage, we construct a directed multigraph G, in which each vertex is a tuple containing an integer mass in the interval $[0, M]$ and a count of the number of each of the symbols in c consumed by a prefix ending at the vertex. Arcs are added between vertices whose mass differs by that of an amino acid and have a compatible count. An arc of G is assigned a cost obtained as a function of the best peaks in T supporting the vertices of the arc. In the second stage, we obtain the N shortest paths in G. Each path must start at the vertex representing mass zero with no symbols from the multiset constraint consumed, and must reach a vertex representing the mass M in which all the symbols appearing in the multiset constraint are consumed.

Intuitively, our dynamic programming algorithm generates a graph with multiple stages where a stage represents a partial set of constraints satisfied so far. More formally, let $V(G)$ denote the vertex set of the directed multigraph G. Let A denote the set of masses of the amino acids represented by the symbols in $\mathcal{A}$. By $span(A)$ we mean the union of the set of numbers that can be written as a sum of elements of A, and the set $\{0\}$. We denote by $\mathcal{A}_c$ the set of symbols $\{a_1, \ldots, a_n\}$ in the constraint c—i.e.,

$c(a_i) > 0$—and by A_c the corresponding masses of the amino acids they represent. Then,

$$V(G) = \left\{ (m, v) : m \in span(A) \quad \text{and} \quad m \leq M; \quad v \in \prod_{i=1}^{n} \{0, \ldots, c(a_i)\} \right\}, \quad (1)$$

where the product is the usual cartesian product of sets. Thus, each vertex (m, v) represents the mass of a prefix weighing m, and n bounded counters, which we denote by $v_1, \ldots, v_n$. The i-th counter keeps a count of the number of a_i symbols consumed by the prefix—ending at that vertex—of any peptide passing through that vertex.

Vertices $x = (m_1, u)$ and $y = (m_2, v)$ in $V(G)$ are related by an arc from x to y if and only if either of the following conditions is satisfied:

1. $m_2 - m_1 \in A \setminus A_c$, and $u = v$, or
2. $m_2 - m_1$ is the mass of $a_i \in \mathcal{A}_c$, and $v_k = \begin{cases} u_k + 1 \text{ if } k = i, \text{ and} \\ u_k, \text{ otherwise.} \end{cases}$

Figure 1 (a) shows a visual representation of the directed multigraph constructed from a small multiset constraint.

We annotate each vertex of the multigraph G with supporting peaks, if any, from the given spectrum. For example, consider the directed multigraph constructed under a constraint $c(\mathsf{C}) = 4$, and consider the vertex $(320, (2))$. This vertex represents a mass of 320 Da, and represents a prefix containing two C out of the minimum of four required, assuming carbamidomethylated Cysteine. We then search the peak list in the spectrum for b-ions (e.g., peaks in the interval $321.00728 \pm \epsilon$ Da) and y-ions (e.g., peaks in the interval $M - 300.98 \pm \epsilon$) supporting this vertex, for a given fragment mass error tolerance of ϵ. After annotating all vertices in this way, we assign costs to each arc in G. In determining the cost of each arc, we use this information about the presence of supporting peaks, their intensity, and the agreement of the mass difference of peaks across an arc with an amino acid mass. Vertices with no support contribute to a penalty for all their arcs. Finally, we attempt to obtain the K least cost paths between the starting vertex of mass zero and a final vertex of mass M and with its prefix symbol counts matching or exceeding the multiset constraint.

More formal details of the algorithm are listed in pseudocode form in Algorithm 1. When $\mathcal{A}_c$ is empty, the algorithm guarantees that every peptide is considered, as is clear from lines 13-19. Line 5 guarantees that no peptide of a mass larger than that reported by the spectrum is considered. Line 26 guarantees that the list of peptides, implied by the list of paths considered, must be of mass M. This argument also holds for unconstrained symbols when $\mathcal{A}_c$ is not empty. When $\mathcal{A}_c$ is not empty, consider any prefix of any peptide $w \in S(c)$. If the prefix contains no constrained letters, then by the arguments above, it is guaranteed to be present as a path in the directed multigraph. If it contains some constrained letters, then their counts and the prefix's mass together must be represented by some vertex in $V(G)$, because of lines 3-12. Finally, only paths ending in a vertex who counts match the multiset constraint and whose mass matches the mass M reported in the spectrum are used for generating peptides. The converse argument proceeds along a similar path. (Note that our algorithm does not generate

Algorithm 1. GENERATING MULTISET-CONSTRAINED DE NOVO CANDIDATES

Require: Constraint $c : \mathcal{A} \to \mathbb{N}, \mathcal{A}_c, A_c$; Spectrum $\mathcal{S} = (T, M)$; Number of candidates K

$V(G) \leftarrow (0, (0, \ldots, 0))$
while more vertices in $V(G)$ remain to be expanded **do**
  $(m, (v_1, \ldots, v_n)) \leftarrow$ next unexpanded vertex from $V(G)$
  for every $a \in \mathcal{A}$ **do**
    if $m + mass(a_i) \leq M$ **then**
      if $a \in \mathcal{A}_c$ **then**
        Let a be the i-th symbol in $\mathcal{A}_c$, denoted by a_i
        if $(m + mass(a_i), (v_1, \ldots, v_i + 1, \ldots, v_n)) \notin V(G)$ **then**
          $(m', v') \leftarrow (m + mass(a_i), (v_1, \ldots, v_i + 1, \ldots, v_n))$
          $V(G) \leftarrow V(G) \cup \{(m', v')\}$
          Mark (m', v') as unexpanded
        end if
      else
        if $(m + mass(a_i), (v_1, \ldots, v_n)) \notin V(G)$ **then**
          $(m', v') \leftarrow (m + mass(a_i), (v_1, \ldots, v_n))$
          $V(G) \leftarrow V(G) \cup \{(m', v')\}$
          Mark (m', v') as unexpanded
        end if
      end if
      Add arc from (m, v) to (m', v')
    end if
  end for
end while
Annotate each vertex with peaks in T corresponding to its mass
Assign weights to each arc
Obtain K shortest paths between $(0, (0, \ldots, 0))$ and $(M, (c(a_1), \ldots, c(a_n)))$
if no such path exists **then**
  Stop and report an unsatisfiable constraint error
else
  Translate each path of vertices into a string over $\mathcal{A}$
  Stop and return this set of peptides
end if

unreachable vertices—for example, $(0, (1))$ in the example above—though we choose to ignore this detail in equation 1 above.)

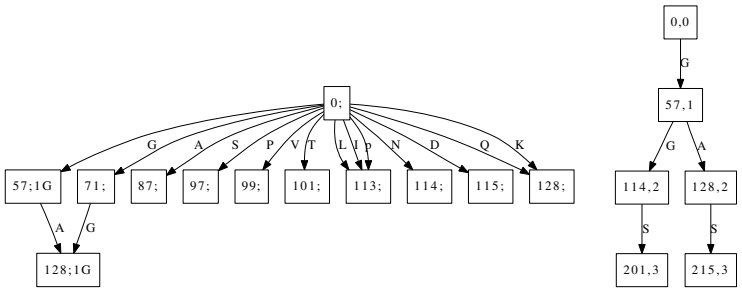
We have omitted several details about some of the steps of our algorithm, such as the arc weighting computation, presence of duplicate and conflated paths, incorporation of terminal modifications, and speed and memory optimizations. While these details may be necessary in an implementation of the algorithm—and our own implementation includes them—they are largely independent of the focus of this paper: demonstrating the feasibility and utility of constrained de novo search. We comment on these details where necessary. A complexity analysis of similar constrained shortest paths problems was carried out by Barrett et al. [4].

3.3 Acyclic Regular Expression Constraint

Let $\mathcal{A}$ be the set of amino acid symbols (including modifications) and let n be a positive integer. An *n-letter acyclic regular expression constraint* is a string $c \in (\mathcal{A} \cup \{\mathcal{A}\})^n$ describing a subset of $\mathcal{A}^\star$, which we denote by $S(c)$. Thus, the string $\mathcal{A}\mathsf{CC}\mathcal{A}\mathcal{A}\mathcal{A}\mathsf{K}\mathcal{A}\mathsf{CC}$ is an example of a 10-letter acyclic regular expression (or regex) constraint. An n-letter regex constraint c has the following interpretation. Every string in $S(c)$ must belong to $\mathcal{A}^n$, and must agree with c at every position, except those containing an $\mathcal{A}$. In the above example of a regex constraint,

$$S(c) = \{w : w \in \mathcal{A}^n \text{ and } w \text{ has } \mathsf{C} \text{ in positions 2,3,9, and 10, and } \mathsf{K} \text{ in position 7}\}$$

For example, GCCPTCKPCC satisfies the regex constraint but CCPCKPCC and AGCCPTCKCC do not.



(a) (b)

Fig. 1. (a) The directed multigraph resulting from the multiset constraint "$c(\mathsf{G}) = 1$" given a spectrum of 128.06 Da. Vertices are labeled with the integer mass of and a count of the constrained symbols in the prefix they represent. In this case, only two paths—GA and AG—satisfy the multiset and mass constraint. (b) The directed multigraph resulting from the regex constraint "$\mathsf{G}\mathcal{A}\mathsf{S}$" given a spectrum of total mass 215.09 Da. Vertices are labeled with the integer mass and the length of the prefix they represent. In this case, the path that satisfies the regex and mass constraint—GAS—is unique.

3.4 An Algorithm for Generating Regex-Constrained Candidates

In its graph and flow, our algorithm for generating regex-constrained peptides is similar to the algorithm for multiset-constrained peptides given above. The main difference is in the information represented in each vertex. Let c be an n-letter regex constraint. In this case,

$$V(G) = \{(m, v) : m \in span(A) \quad \text{and} \quad m \leq M; \quad v \in \{0, \ldots, n\}\}, \tag{2}$$

Each vertex represents the mass of a prefix of every path passing through it, and a count of the number of letters in the prefix. Vertices $x = (m_1, v)$ and $y = (m_2, v+1)$ in $V(G)$ are related by an arc from x to y if and only if $m_2 - m_1 \in A$. Other details are similar to the multiset algorithm described above; the differences are formally presented in Algorithm 2 below. Figure 1 (b) shows a visual representation of the directed multigraph constructed for a small regex constraint.

3.5 Constrained Spectral Search and Clustering

Given a spectrum, in many cases, *de novo* sequencing of the complete peptide may be difficult. Typically, this is due to the quality of the spectrum, unavailability of the complete ladder of peaks in any single spectrum due to digestion, or low mass accuracy of the fragments or the precursor. In such instances, it is desirable to have tool that can quickly search for other spectra that describe the unknown peptide under consideration. We implemented a simple spectral search tool for this purpose, similar to SALSA [24]. Given a spectrum, we consider the set of its peaks as vertices and construct a directed multigraph G in which we add an arc between any two peaks separated by the mass of some amino acid, including modified amino acids. Then, we enumerate all maximal distinct paths in G. This results in a list of short peptide fragments, not necessarily of the mass reported in the spectrum, all of which are supported by peaks in the given spectrum. This "spectral fingerprint" can be used to search for spectra containing peaks supporting a particular peptide fragment. In the context of conotoxin spectra, we have found this tool to be useful for filtering out spectra that contain a "CC fingerprint" and thus, are likely to be sequenceable conotoxins. It can also be used for clustering spectra.

Algorithm 2. GENERATING REGEX-CONSTRAINED DE NOVO CANDIDATES

Require: Constraint $c : \{1, \ldots, n\} \rightarrow \mathcal{A}$; Spectrum $\mathcal{S} = (T, M)$; Number of candidates K

1. $V(G) \leftarrow (0, 0)$
2. **while** more vertices in $V(G)$ remain to be expanded **do**
3. $(m, i) \leftarrow$ next unexpanded vertex from $V(G)$
4. **if** $i = n$ **then**
5. Go to line 23
6. **end if**
7. **if** $c(i + 1) =$ "$\mathcal{A}$" **then**
8. $\mathcal{B} \leftarrow \mathcal{A}$
9. **else**
10. $\mathcal{B} \leftarrow \{c(i + 1)\}$
11. **end if**
12. **for** every $a \in \mathcal{B}$ **do**
13. **if** $m + mass(a) \leq M$ **then**
14. **if** $(m + mass(a), i + 1) \notin V(G)$ **then**
15. $(m', i') \leftarrow (m + mass(a), i + 1)$
16. $V(G) \leftarrow V(G) \cup \{(m', i')\}$
17. Mark (m', i') as unexpanded
18. **end if**
19. Add arc from (m, i) to (m', i')
20. **end if**
21. **end for**
22. **end while**
23. (Same as lines 24-25 in Algorithm 1 above)
24. Obtain K shortest paths between $(0, 0)$ and (M, n)
25. (Same as lines 27-32 in Algorithm 1 above)

In addition to the above algorithms, we have also implemented algorithms that allow ordered multiset-constraints (e.g., two C followed, not necessarily immediately, by a W), combine multiset and regex constraints (e.g. GCCKP followed by two C), and impose mass intervals in which a constraint must be satisfied. We postpone discussion of these algorithms to future work.

4 Application to Conotoxins

We obtained MS/MS data of *Conus textile* venom from Brian Chait's laboratory at Rockefeller University and of *C. textile* and *C. stercusmuscarum* venom from John Yates's laboratory at the Scripps Research Institute. The Rockefeller data [37] were LTQ MS/MS spectra, both CID and ETD, with low-accuracy precursor and fragment masses. We did not obtain Rockefeller's charge-enhanced precursor data [37], only the standard carbamidomethylated cysteine. Sample preparations and data acquisition strategies were as described previously [37,36]. The Scripps data [36] were HCD and CID Orbitrap MS/MS spectra with high-accuracy precursor and fragment masses.

Both *C. textile* data sets had been analyzed previously by database search and, in the case of the Rockefeller data, a limited amount of manual de novo sequencing. *C. textile* is one of the better studied cone snails, with a large amount of venom, and GenBank contains about 100 (redundant) *C. textile* entries, more than half of which are putative toxins. One of the goals in proteomic analysis is to observe the toxins in their mature forms, meaning with the post-translational modifications and exact termini. Conotoxins are heavily modified peptides of lengths about 10–40 residues, and known modifications include bromotryptophan, hydroxyproline, hydroxyvaline, oxidized methionine, and amidated C-terminus. Both the Rockefeller and Scripps studies claimed 31 *C. textile* toxins observed in their final form. The venom contains about 90 toxins, as estimated by the number of disulfide-bonded precursors [37]. For *C. stercusmuscarum*, there is very little sequence data available, only seven GenBank entries, none of which are annotated as toxins, so this data was essentially unanalyzed when we received it.

4.1 Validation on Known Sequences

We implemented the algorithms for constrained *de novo* search listed above into a single command-based interactive tool which we call CONOVO. The tool is capable of reading in a set of CID or ETD spectra and accepting a sequence of commands to operate on them. These include commands for de-isotoping and de-charging spectra, adding, deleting and ignoring peaks in spectra, normalizing peak intensities in a spectrum, generating a spectral fingerprint, collecting the top peaks from several spectra into a single spectrum, constructing, examining, and modifying directed multigraphs for *de novo* search, and generating candidate peptides. We wrote scripts for processing all the spectra that we received from the Yates and Chait laboratories. Our scripts processed each spectrum by issuing commands to CONOVO to load, deisotope and decharge the spectrum, and then generate candidates under various constraints, or without any constraints. After candidate generation in each case, the candidates were scored by the ByOnic scorer [6] and the highest scoring candidate was logged along with a detailed report explaining the score. We pointed our scripts to the spectra and executed the scripts without any subsequent human intervention.

We first report results from 79 *C. textile* CID spectra from both laboratories. These spectra describe cysteine-rich conotoxins whose complete and correct sequences are known. Our scripts sequenced these spectra using purely multiset-constrained and purely regex-constrained *de novo* search. We also ran an unconstrained *de novo* search under the same conditions as the multiset-constrained *de novo* search. The answer found under all three conditions on these spectra agreed with the correct answer, modulo K-Q, I-L, M-F, and GG-N substitutions, if any. Our scripts executed all of the following multiset-constraints on each spectrum: $c(\mathsf{C}) = 2$, $c(\mathsf{C}) = 3$, . . ., $c(\mathsf{C}) = 6$. We obtained regex constraints directly from the correct answer by retaining all C and substituting all other letters with $\mathcal{A}$ in the correct answer.

Figure 2 shows the decimal logarithm of the position of the correct answer in the generated candidate list in the constrained case (X-axis) and the unconstrained (Y-axis) case. The left plot shows the multiset-constrained case and the right plot shows the regex-constrained case. Points near or on the diagonal $x = y$ line result from spectra on which both the constrained search and the unconstrained search produced the right answer at the same or similar position in their respective candidates list. Points below the diagonal result from spectra on which the constrained search produced the correct candidate at a position in its candidates list that was an order of magnitude lower than the position of the correct candidate on the unconstrained candidates list. For example, the point $(4.47, 2.03)$ in the multiset case (left) corresponds to the correct *C. textile* conotoxin scaffold precursor SCCNAGFCRFGCTPCCY, which was generated at position 29,610 by the unconstrained search and at positions 109 and 1 under the $c(\mathsf{C}) = 6$ and the $\mathcal{A}\mathsf{CC}\mathcal{A}\mathcal{A}\mathcal{A}\mathcal{A}\mathsf{C}\mathcal{A}\mathcal{A}\mathcal{A}\mathsf{C}\mathcal{A}\mathcal{A}\mathsf{CC}\mathcal{A}$ constraints. The plot confirms our hypothesis that constraints can be extremely effective in improving the efficiency of the *de novo* search by reducing the search space to a subset where all the candidates satisfy *a priori* knowledge. We note that a regex constraint is more effective than a multiset constraint, but it requires much stronger *a priori* knowledge: one must supply the exact position of every letter in the constraint.

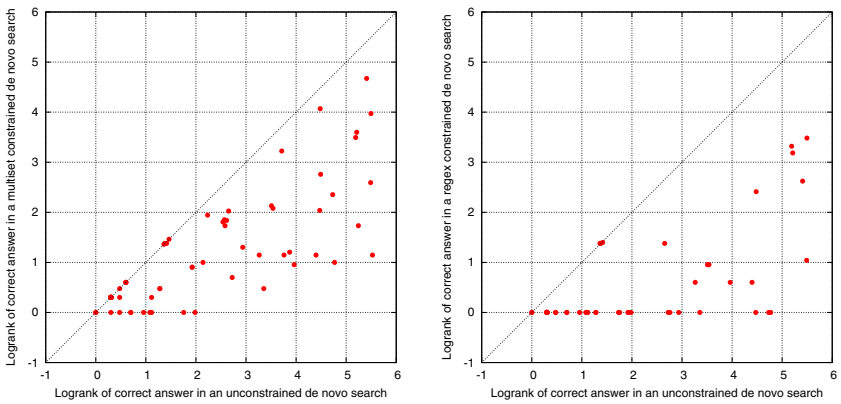


Fig. 2. A comparison of a multiset-constrained (left) or regex-constrained (right) *de novo* search with an unconstrained *de novo* search for peptide candidates for 79 cysteine-rich conotoxin spectra of the venom of *C. textile*

4.2 Discovery of Novel Conotoxins

In addition to known conotoxins, CONOVO also found peptides that appear to be new. In this paper, we report two sequences that were, to the best of our knowledge, unknown; we will report all other sequences once they have been verified by synthesis.

Figures 3(a) and 3(b) show the spectra describing the novel conotoxin found in the *C. textile* venom and the *C. stercusmuscarum* venom respectively. Figures 3(c) and 3(d) show spectra describing a prefix and a suffix of the novel *C. stercusmuscarum* toxin. In the *C. Textile* data, CONOVO found the sequence

C[+57]C[+57]GP[+16]TAC[+57]LAGC[+57]KPC[+57]C[+57][-1]

in at least 16 spectra of mass 1786 Da and 1802 Da. The 1802 Da spectra indicate a PTM on the second proline. The identifications were obtained from the multiset-constrained search described in Section 4.1. Figure 3(a) shows one of the spectra describing this novel conotoxin.

In the *C. stercusmuscarum* data, CONOVO found the following sequence:

APAC[+57]C[+57]GPGASC[+57]PRYFKDNFLC[+57]GC[+57]C[+57]

The prefix APACCGPGASCPR, but with a few incorrect letters, was found in the multiset-constrained search described above. After the constrained search completed, we collected potentially related spectra with a spectral fingerprint search described above.

We sequenced the spectrum in Figure 3(c) and obtained APACCGPGASCPRYF, which, due to its odd number of C, we guessed was a prefix of the complete sequence of the toxin. We then found a spectrum for the complete toxin (Figure 3(b)) using a wild-card with mass up to 2000 Da. (Notice that this spectrum would be hard to find by spectrum similarity to Figure 3(a).) This spectrum is not sequenceable on its own. We then found the spectrum shown in Figure 3(d) using spectral search.

5 Discussion

Constrained de novo sequencing is a new peptide identification approach that is especially well suited to studies focused on diverse but homologous protein families such as conotoxins or antibodies. We found the approach advantageous in the conotoxin studies described here, and by the end of the project, our data-analysis approach was fully automated, with human intervention required only to inspect ByOnic's scoring reports, and reconcile spectra that clearly contained closely related or identical peptides, but had incompatible top-scoring sequences.

Constrained sequencing is advantageous in different ways. First, in the case of sequenceable spectra with complete fragmentation, constraints reduce the space of plausible candidates while incorporating specific expert knowledge, thus boosting the efficacy of the scorer by eliminating spurious candidates. Such a reduction in candidate space could in principle be achieved by running an unconstrained search followed by an elimination step imposing the constraints, but for long peptides, this is not feasible. If the size of the generated candidates list is limited—as is the case in practice for long peptides—then the correct answer may not even be generated by an unconstrained search, rendering the elimination step ineffective. In this case, a constrained search is

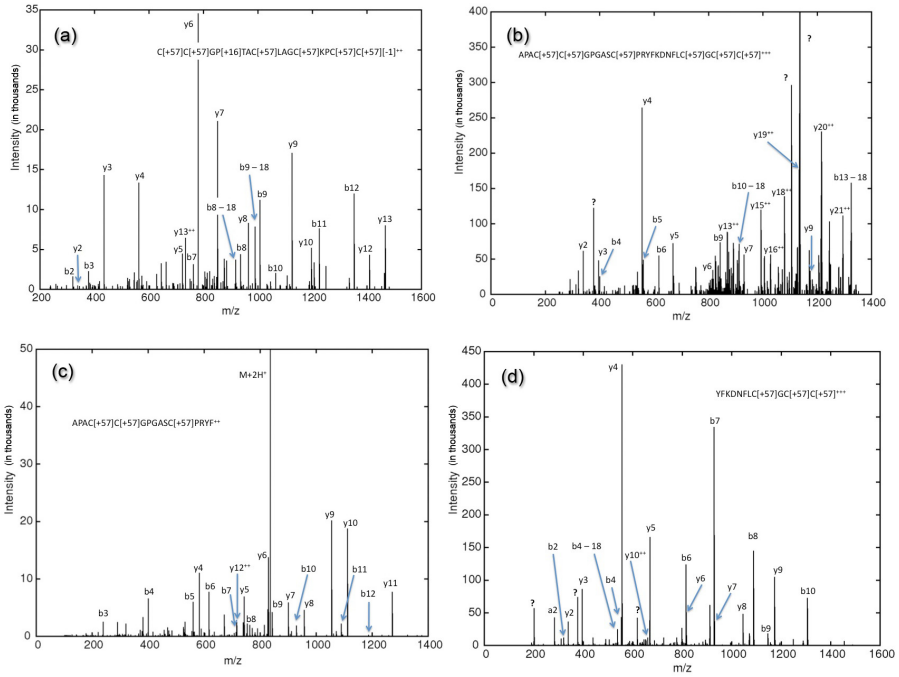


Fig. 3. (a) A novel *C. textile* toxin. All mass errors are less than 4 ppm. Despite the high mass accuracy, this spectrum would be quite challenging to sequence without some prior knowledge, because of the two PTMs (hydroxyproline and amidated C-terminus) and the missing cleavages at b1/y14 and b4/y11 (after hydroxyproline). The closest known conotoxin is CCGPTACMAGCR-PCC, two substitutions away. (b) A novel *C. stercusmuscarum* toxin with no BLAST hits in GenBank with E-value below 1.0. This spectrum gives what we believe is the complete toxin, observed in the undigested venom. Mass errors are less than 10 ppm, but with software recalibration of the m/z readings, the errors can be reduced to less than 4 ppm. (c) The N-terminal half of the novel *C. Stercus muscarum* toxin, also observed in the undigested venom. All errors are less than 4 ppm. (d) The C-terminal half of the novel *C. Stercus muscarum* toxin, observed in a tryptic digest of the venom. With software recalibration, all errors are less than 4 ppm.

a natural and effective solution. The reduction in the size of the candidate space may span orders of magnitude as revealed by a simple counting argument (see Figure 4) and illustrated by our experiments.

Second, constraints can actually bridge missing cleavages and sequence otherwise unsequenceable spectra. Consider a spectrum containing all the peaks supporting any candidate of the form PEPTIDE$\mathcal{A}'\mathcal{A}'\mathcal{A}'\mathcal{A}'$ where $\mathcal{A}'$ does not contain C, and all the peaks supporting PEPTIDECCCC. It is plausible that a scorer may rank candidates from both sets equally or even prefer the former candidates over the latter as a result of peak position noise. Yet, even a simple regex constraint like PEPTIDEC$\mathcal{AAA}$ would be sufficient for the scorer to rule out the former set in favor of the latter candidate. In such a case, the gain from a single-letter regex constraint is more significant than the reduction in space provided by fixing a single letter.

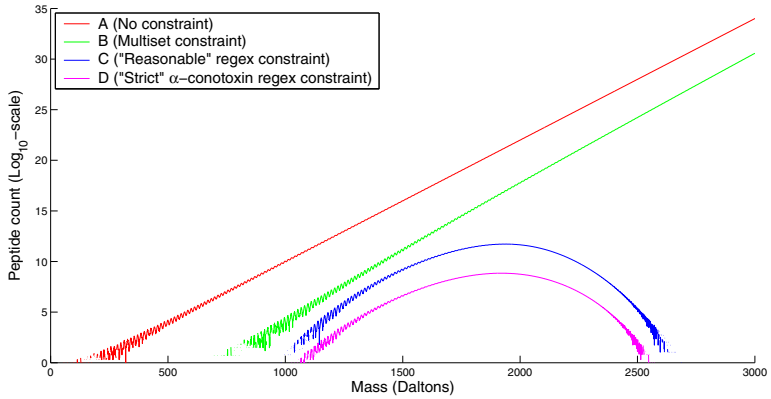


Fig. 4. Size of the candidate space: A without constraints; B with multiset constraints; C with regex constraints like $\mathsf{CC}\mathcal{A}^{3\text{ or }4}\mathsf{C}\mathcal{A}^{3\text{ to }7}\mathsf{C}$; and D with an α-conotoxin regex constraint [39]

Nevertheless, constraints are no panacea for *de novo* sequencing. While the multiset-constrained candidate generation succeeded in sequencing a majority of the known sequences and discovering some unknown sequences, it was not successful on all spectra. In most such cases, we were able to obtain a candidate within an edit distance of three or four of the correct answer. We were then able to complete the sequence either manually, or by running a search on a database of newly discovered sequences, or by using the letters found so far, as a regex constraint. We used ByOnic's wild-card feature [8,9] to be useful in computing missing masses in incomplete *de novo* sequences, which we were later able to fill using constrained or unconstrained de novo search.

For lower accuracy spectra, there were instances where our *de novo* search produced several plausible candidates. We discovered a heuristic for separating false positives in such instances. We checked whether the mass errors of the b- and y-ions matched in magnitude. We rejected candidates, for example, in which y-ions had mass errors of 50 ppm which b-ions had mass errors of 5 ppm.

We also found that constraints were not very useful on spectra of very high or low quality, since the former were readily sequenceable without constraints while the latter were mostly unsequenceable. We also note that incorrect constraints—e.g., requiring a letter that is absent in the correct sequence—can ruin a *de novo* search. The user should start with a weak constraint and gradually strengthen it as more of the sequence becomes known. Thus, several iterations of a combination of multiset and regex constraints and ByOnic's wild card may prove to be an effective *de novo* protocol.

We found spectral search to be a handy tool for gathering spectral evidence for low confidence letters and for ruling out competing candidates. While spectral clustering would have been helpful, we did not use any clustering in this project.

Acknowledgments

This work was supported by NIGMS grant R21 GM094557, an ARRA supplemental to grant R21 GM085718, and a NSF Computing Innovations Fellowship.

References

1. Bandeira, N., Tsur, D., Frank, A., Pevzner, P.A.: Protein identification by spectral networks analysis. Proc. Natl. Acad. Sci. USA 104, 6140–6145 (2007)
2. Bandeira, N., Clauser, K.R., Pevzner, P.A.: Assembly of peptide tandem mass spectra from mixtures of modified proteins. Molecular Cell. Proteomics 6, 1123–1134 (2007)
3. Bandeira, N., Pham, V., Pevzner, P., Arnott, D., Lill, J.R.: Automated de novo protein sequencing of monoclonal antibodies. Nature Biotechnology 26, 1336–1338 (2008)
4. Barrett, C., Jacob, R., Marathe, M.: Formal language constrained path problems. SIAM J. on Computing 30, 809–837 (2000)
5. Bartels, C.: Fast algorithm for peptide sequencing by mass spectrometry. Biomedical and Environmental Mass Spectrometry 19, 363–368 (1990)
6. Bern, M., Cai, Y., Goldberg, D.: Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal. Chem. 79, 1393–1400 (2007)
7. Bern, M., Goldberg, D.: De novo analysis of peptide tandem mass spectra by spectral graph partitioning. J. Computational Biology 13, 364–378 (2006)
8. Bern, M., Phinney, B.S., Goldberg, D.: Reanalysis of Tyrannosaurus rex Mass Spectra. J. Proteome Res. 8, 4328–4332 (2009)
9. Bern, M., Saladino, J., Sharp, J.S.: Conversion of methionine into homocysteic acid in heavily oxidized proteomics samples. Rapid Commun. Mass Spectrom. 24, 768–772 (2010)
10. Chen, T., Kao, M.-Y., Tepel, M., Rush, J., Church, G.M.: A dynamic programming approach to de novo peptide sequencing by mass spectrometry. J. Computational Biology 8, 325–337 (2001)
11. Dančik, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A.: De novo peptide sequencing via tandem mass spectrometry. J. Computational Biology 6, 327–342 (1999)
12. Datta, R., Bern, M.: Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. J. Comput. Biol. 16, 1169–1182 (2009)
13. Depontieu, F.R., Qian, J., Zarling, A.L., McMiller, T.L., Salay, T.M., Norris, A., English, A.M., Shabanowitz, J., Engelhard, V.H., Hunt, D.F., Topalian, S.L.: Identification of tumor-associated, MHC class II-restricted phosphopeptides as targets for immunotherapy. Proc. Natl. Acad. Sci. USA 106, 12073–12078 (2009)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience, Hoboken (2000)
15. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., Gygi, S.P.: Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nature Biotechnology 22, 214–219 (2004)
16. Eng, J.K., McCormack, A.L., Yates III., J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989 (1994)
17. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., Buhmann, J.M.: NovoHMM: A hidden Markov model for de novo peptide sequencing. Anal. Chem. 77, 7265–7273 (2005)
18. Eppstein, D.: Finding the k shortest paths. SIAM J. Computing 28, 652–673 (1998)
19. Frank, A., Pevzner, P.: PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Anal. Chem. 77, 964–973 (2005)
20. Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., Pevzner, P.A.: De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. J. Proteome Research 6, 114–123 (2007)

21. Graehl, J.: Implementation of David Eppstein's k Shortest Paths Algorithm, http://www.ics.uci.edu/~eppstein/
22. Havilio, M., Haddad, Y., Smilansky, Z.: Intensity-based statistical scorer for tandem mass spectrometry. Anal. Chem. 75, 435–444 (2003)
23. Kaas, Q., Westermann, J.C., Halai, R., Wang, C.K., Crak, D.J.: ConoServer, a database for conopeptide sequences and structures. Bioinformatics 445, 445–446 (2008)
24. Liebler, D.C., Hansen, B.T., Davey, S.W., Tiscareno, L., Mason, D.E.: Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. Anal. Chem. 74, 203–210 (2002)
25. Liu, X., Han, Y., Yuen, D., Ma, B.: Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. Bioinformatics 25, 2174–2180 (2009)
26. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Comm. in Mass Spectrometry 17, 2337–2342 (2003), http://www.bioinformaticssolutions.com
27. Nair, S.S., Nilsson, C.L., Emmett, M.R., Schaub, T.M., Gowd, K.H., Thakur, S.S., Krishnan, K.S., Balaram, P., Marshall, A.G.: De novo sequencing and disulfide mapping of a bromotryptophan-containing conotoxin by Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. 78, 8082–8088 (2006)
28. Pham, V., Henzel, W.J., Arnott, D., Hymowitz, S., Sandoval, W.N., Truong, B.-T., Lowman, H., Lill, J.R.: De novo proteomic sequencing of a monoclonal antibody raised against OX40 ligand. Analytical Biochemistry 352, 77–86 (2006)
29. Resemann, A., Wunderlich, D., Rothbauer, U., Warscheid, B., Leonhardt, H., Fuschser, J., Kuhlmann, K., Suckau, D.: Top-Down de Novo Protein Sequencing of a 13.6 kDa Camelid Single Heavy Chain Antibody by Matrix-Assisted Laser Desorption Ionization-Time-of-Flight/Time-of-Flight Mass Spectrometry. Anal. Chem. 82, 3283–3292 (2010)
30. Savitski, M.M., Nielsen, M.L., Kjeldsen, F., Zubarev, R.A.: Proteomics-Grade de Novo Sequencing Approach. J. Proteome Research, 2348–2354 (2005)
31. Shevchenko, A., et al.: Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal. Chem. 73, 1917–1926 (2001)
32. Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J., Hunt, D.F.: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. USA 101, 9528–9533 (2004)
33. Tabb, D.L., Smith, L.L., Breci, L.A., Wysocki, V.H., Lin, D., Yates III., J.R.: Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic digests. Anal. Chem. 75, 1155–1163 (2003)
34. Tabb, D.L., MacCoss, M.J., Wu, C.C., Anderson, S.D., Yates III., J.R.: Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. Anal. Chem. 75, 2470–2477 (2003)
35. Taylor, J.A., Johnson, R.S.: Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal. Chem. 73, 2594–2604 (2001)
36. Tayo, L.L., Lu, B., Cruz, L.J., Yates III., J.R.: Proteomic analysis provides insights on venom processing in *Conus textile*. J. Proteome Research 9, 2292–2301 (2010)
37. Ueberheide, B.M., Fenyö, D., Alewood, P.F., Chait, B.T.: Rapid sensitive analysis of cysteine rich peptide venom components. Proc. Natl. Acad. Sci. USA 106, 6910–6915 (2009)
38. Zhang, Z., McElvain, J.S.: De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. Anal. Chem. 72, 2337–2350 (2000)
39. Alpha-conotoxin family signature. Accession number PS60014, ProSite ExPASy Proteomics Server (March 2005)