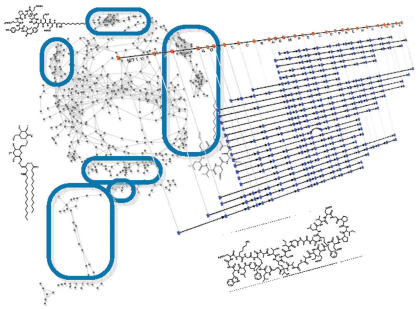
Molecular BioSystems c2mb25085c



The spectral networks paradigm in high throughput mass spectrometry Q1

Adrian Guthals, Jeramie D. Watrous, Pieter C. Dorrestein and Nuno Bandeira*

We discuss the current state of spectral networks algorithms and discuss possible future directions for automated interpretation of spectra from any class of molecules. Q2

Please check this proof carefully. **Our staff will not read it in detail after you have returned it.** Translation errors between word-processor files and typesetting systems can occur so the whole proof needs to be read. Please pay particular attention to: tabulated material; equations; numerical data; figures and graphics; and references. If you have not already indicated the corresponding author(s) please mark their name(s) with an asterisk. Please e-mail a list of corrections or the PDF with electronic notes attached – do not change the text within the PDF file or send a revised manuscript.

Please bear in mind that minor layout improvements, e.g. in line breaking, table widths and graphic placement, are routinely applied to the final version.

Please note that, in the typefaces we use, an italic vee looks like this: v, and a Greek nu looks like this: ν.

We will publish articles on the web as soon as possible after receiving your corrections; **no late corrections will be made.**

Please return your **final** corrections, where possible within **48 hours** of receipt, by e-mail to: molbiosyst@rsc.org

Reprints—Electronic (PDF) reprints will be provided free of charge to the corresponding author. Enquiries about purchasing paper reprints should be addressed via: http://www.rsc.org/publishing/journals/guidelines/paperreprints/. Costs for reprints are below:

Reprint costs		
No of pages	Cost (per 50 copies)	
	First	Each additional
2-4	£225	£125
5-8	£350	£240
9-20	£675	£550
21-40	£1250	£975
>40	£1850	£1550
Cost for including cover of journal issue: £55 per 50 copies		

Queries are marked on your proof like this Q1, Q2, etc. and for your convenience line numbers are indicated like this 5, 10, 15, ...

Query reference	Query	Remarks
Q1	[INFO-1] Please carefully check the spelling of all author names. This is important for the correct indexing and future citation of your article. No late corrections can be made. FOR YOUR INFORMATION: You can cite this paper before the page numbers are assigned with: (authors), Mol. BioSyst., DOI: 10.1039/c2mb25085c.	
Q2	Is the contents entry text acceptable ? If not, please provide an alternative which should be no longer than 30 words.	
Q3	Please indicate where Fig. 3 should be cited in the text.	
Q4	Ref. 29: Can this reference be updated yet? Please supply details to allow readers to access the reference (for references where page numbers are not yet known, please supply the DOI).	
Q5	Please provide an author surname for the 6th author in reference 32.	
Q6	Ref. 51 and 52: Please provide the following details: full list of author names (including initials).	
Q7	Ref. 78, 101, 102 and 104: Can these references be updated yet? Please supply details to allow readers to access the reference (for references where page numbers are not yet known, please supply the DOI).	

Molecular BioSystems

Dynamic Article Links

Cite this: DOI: 10.1039/c2mb25085c

www.rsc.org/molecularbiosystems

PAPER

The spectral networks paradigm in high throughput mass spectrometry†

Adrian Guthals,[a] Jeramie D. Watrous,[bc] Pieter C. Dorrestein[bc] and Nuno Bandeira*[ac]

Received 11th March 2012, Accepted 20th April 2012
DOI: 10.1039/c2mb25085c

High-throughput proteomics is made possible by a combination of modern mass spectrometry instruments capable of generating many millions of tandem mass (MS^2) spectra on a daily basis and the increasingly sophisticated associated software for their automated identification. Despite the growing accumulation of collections of identified spectra and the regular generation of MS^2 data from related peptides, the mainstream approach for peptide identification is still the nearly two decades old approach of matching one MS^2 spectrum at a time against a database of protein sequences. Moreover, database search tools overwhelmingly continue to require that users guess in advance a small set of 4–6 post-translational modifications that may be present in their data in order to avoid incurring substantial false positive and negative rates. The spectral networks paradigm for analysis of MS^2 spectra differs from the mainstream database search paradigm in three fundamental ways. First, spectral networks are based on matching spectra against other spectra instead of against protein sequences. Second, spectral networks find spectra from related peptides even before considering their possible identifications. Third, spectral networks determine consensus identifications from *sets* of spectra from related peptides instead of separately attempting to identify one spectrum at a time. Even though spectral networks algorithms are still in their infancy, they have already delivered the longest and most accurate *de novo* sequences to date, revealed a new route for the discovery of unexpected post-translational modifications and highly-modified peptides, enabled automated sequencing of cyclic non-ribosomal peptides with unknown amino acids and are now defining a novel approach for mapping the entire molecular output of biological systems that is suitable for analysis with tandem mass spectrometry. Here we review the current state of spectral networks algorithms and discuss possible future directions for automated interpretation of spectra from any class of molecules.

1 Introduction

The success of tandem mass spectrometry (MS^2) approaches to peptide identification is partly due to advances in computational techniques allowing for the reliable interpretation of MS^2 spectra. Mainstream computational techniques mainly fall into two categories: database search approaches that score each spectrum against peptides in a sequence database[1–4] and *de novo* techniques that directly reconstruct the peptide sequence from each spectrum.[5–8] The combination of these methods with advances in high throughput MS^2 have promoted accelerated growth of spectral libraries–collections of peptide MS^2 spectra whose identifications were validated by accepted statistical methods[9,10] and often also manually confirmed by mass spectrometry experts. A similar concept of spectral archives was also recently proposed to denote spectral libraries including "interesting" non-identified spectra[11] (*i.e.* unidentified recurring spectra with good *de novo* reconstructions). The growing availability of these large collections of MS^2 spectra has reignited the development of alternative peptide identification approaches based on spectral matching[12–14] and alignment[15–17] algorithms.

The dominant paradigm for high-throughput protein identification is based on trypsin digestion of extracted proteins to produce peptides followed by tandem mass spectrometry to generate single-peptide MS^2 spectra that are then computationally matched one spectrum at a time against protein sequence databases to finally obtain peptide and protein identifications. This paradigm has been the basis of nearly all large-scale proteomics studies to date despite its typical low spectrum identification rate of only 15–30% because enzymatic digestion generates multiple peptides per protein and, in the extreme, only one peptide needs to be

[a] *Dept. Computer Science and Engineering, University of California, San Diego, USA. E-mail: bandeira@ucsd.edu*
[b] *Department of Pharmacology and Department of Chemistry and Biochemistry, University of California, San Diego, USA*
[c] *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, USA*

† Published as part of a themed issue dedicated to Emerging Investigators.

identified per protein (though more are usually preferred) to enable protein-level quantification and comparison across multiple tissues or experimental conditions. However, the serious downside of this low identification rate is that it consistently leads to missing information on non-tryptic peptides and yields very low protein sequence coverage, thus substantially limiting the chances of detecting alternative splicing or to identify and localize post-translational modifications (PTMs). In fact, the limitations of PTM search are so dire that most labs still only allow for 4–6 PTMs per search (about half or which due to sample handling procedures) even though more than 500 PTMs are known and listed in UniMOD.

Peptidomics, defined as the study of endogenous peptides, is an abundant source of drug candidates derived from neuropeptides,[18] toxins[19] and non-linear cyclic peptides.[20] Conversely, endogenous peptides are also valuable as therapeutic targets[21] (neuropeptides) and antigenic peptides are key in immunotherapeutic strategies[22] (MHC class-I/II peptides). Despite its critical importance, peptidomics research continues to suffer from the inadequate reutilization of computational tools primarily developed for proteomics since (a) endogenous peptides are not suitable for enzymatic digestion (as it eliminates the active peptide form), (b) tend to be modified with unexpected PTMs, (c) often contain sequence polymorphisms and (d) generally lack the ''MS-friendly'' features of trypsin-digested peptides. As such, each endogenous peptide must be identified ''on its own'' (not being able to benefit from multiple peptides per protein as in proteomics) and new identification algorithms are needed to be able to handle non-tryptic peptides of atypical lengths[21] (*e.g.*, $\leq$6 AA or $\geq$ 35 AA) containing unexpected PTMs, sequence polymorphisms[20,23] and often featuring non-linear structures.[19,20] Finally, Metaproteomics analysis of environmental samples from host-pathogen interactions[24] and microbial communities (as in the Human Microbiome Project) requires the ability to search mass spectrometry data against very large databases and, in many cases, against six-frame translations of poorly-annotated genomes or even just assembled DNA reads. This enormous growth in the size of the sequences database and the need to allow for polymorphisms and/or unexpected PTMs results in a combined search space so large that 90–95% of all spectra are commonly discarded as unidentified, thus severely limiting proteomics analysis of the role of microbiomes in health and disease.[25]

We argue that overcoming the identification bottleneck will require new ways of thinking about MS^2 spectra in order to develop new ways of interpreting them. In particular, we describe how the spectral networks paradigm differs from the current mainstream paradigm and illustrate its potential with applications where current paradigms perform poorly or completely fail. By finding spectra from related peptides even before considering their possible identifications and using these spectra to determine consensus identifications from sets of spectra from related peptides instead of separately attempting to identify one spectrum at a time, the spectral networking paradigm is capable of addressing many of the pitfalls of mainstream spectra identification paradigms. In addition to improving identification by significantly increasing signal-to-noise ratios and deconvoluting MS^2 ion types, spectral networks further open up new computational avenues for analysis of natural products and non-peptidic molecules, including compounds with non-linear structures, novel amino acids or post-translational modifications, lipids, glycans and other families of compounds.

2 Spectral library matching

The repeated acquisition and reliable identification of MS^2 spectra from a range of biological systems including various microbial species, mammalian tissues and cell lines has led to the accumulation of large collections of identified MS^2 spectra from mostly unmodified or partially modified peptide sequences. As a result, peptide identification by matching spectra of unidentified peptides against *spectral libraries* of identified peptide spectra[14] has recently gained new relevance, especially since the introduction of decoy spectral libraries[26] for calculation of false discovery rates.[10,27] Searching against libraries of predicted spectra is also a promising emerging approach.[28,29]

The potential of spectral libraries to improve peptide identification is well illustrated by the recent example of the NeuroPedia[30] spectral library of identified neuropeptide spectra. Neuropeptides are peptide neurotransmitters and hormones that mediate cell-to-cell communication for regulation of physiological functions and biological processes.[31] Understanding the role and regulation of neuropeptide forms in health, disease, and drug treatments requires the ability to globally analyze neuropeptide expression in an unbiased form. Mass spectrometry based neuropeptidomics is highly suited for untargeted, global neuropeptides studies.[31–35] However, the unique characteristics of neuropeptides (*i.e.* short/long sequences or non-tryptic) presents difficulties for identification from tandem mass spectrometry with traditional database search tools. For example, short neuropeptides can lead to inaccurate search results as database search tools usually assign lower scores to short peptides. Conversely, long or non-tryptic neuropeptides are difficult to identify since database search tools are trained for tryptic peptides cleaved at K/R and because peptide fragmentation processes for long neuropeptides is usually not efficient. In addition, as current databases mature, querying the larger search space requires more time due to the increase in the number of comparisons which ultimately reduces the number of identifications by allowing a higher probability for false positive matches.[27] Since many spectral libraries, such as NeuroPedia, are directly searchable using mass spectrometry data, the caveats associated with matching experimental data against MS^2 spectra predicted from a protein sequence no longer apply as irregularities in fragmentation efficiency will be shared amongst the annotated and unannotated spectra. In addition to the expected improvement in sensitivity from searching against a small targeted sequence database, the neuropeptide spectral libraries further improve identification efficiency, sensitivity and reliability by considering all spectral features, including actual fragment intensities, neutral losses from fragments, and various uncommon or even unknown fragments to determine the best matches. As such, NeuroPedia was shown to improve peptide identification by up to ten fold (at the same false

discovery rate[10,26,27] but searching against a much smaller space of possible matches).

In addition to improving peptide identification, spectral library search opens up new possibilities for interpretation of MS^2 spectra. For example, mainstream approaches were developed under the ubiquitous assumption that each MS^2 spectrum is generated from a single peptide. While chromatographic procedures greatly contribute to making this a reasonable assumption, there are several situations where it is difficult or even impossible to separate pairs of peptides. Examples include certain permutations of the peptide sequence or post-translational modifications (PTMs, see[36] for examples of co-eluting histone modification variants). In addition, innovative experimental setups have demonstrated the potential for increased throughput in peptide identification using mixture spectra–examples include Data-Independent Acquisition[37] Ion-Mobility Mass Spectrometry[38] and MSE strategies.[39] To address the resulting computational bottleneck, we introduced the first spectral library-based approach (M-SPLIT[40]) for identification of mixture spectra generated from more than one peptide. Theoretical bounds were proposed to prune the search space using branch-and-bound techniques and further improved using a new projected-cosine metric. In brief, M-SPLIT uses single-peptide matches to prune the search space for mixture peptides–it first matches experimental spectra to single-peptide spectra and then attempts to improve the score of the match by adding more single-peptide matches to form mixture-spectrum matches (false discovery rates also controlled using decoy spectral libraries[26]). Thus, M-SPLIT dramatically reduces the search space by six orders of magnitude and is able to deliver results at an average of 2 s/spectrum (on a regular laptop with a Pentium Core2Duo, 1.6Ghz, 2Gb RAM), even when searching against proteome-scale spectral libraries. Despite considering only a tiny fraction of the whole search space, benchmarks on both simulated and experimental data consistently show that M-SPLIT[40] has both high sensitivity (≈94%) and high accuracy (up to ≈98%).

3 Spectral networks of unidentified spectra

The simplest example of a spectral network is the detection of MS^2 spectra from repeated acquisition of the same peptide in the same or multiple mass spectrometry runs; in these cases every node in the network is an individual spectrum in a separate run and edges between nodes indicate that the connected nodes represent spectra from the same peptide. Typically, in MS^2 analysis, each mass spectrum in the data set is searched against a sequence database. At times, this can be very inefficient since MS^2 data sets contain many redundancies (it is common for peptides to get selected for fragmentation more than once[41]). When mass spectra are collected from several runs, such redundancies can add up to hundreds and even thousands of spectra from the same peptide. Instead of repeating the identification process for each spectrum, it can be beneficial to perform the search only once using a representative consensus spectrum per peptide and later apply the results to all similar spectra.[41–43] By analyzing only representative spectra (one per cluster of spectra from each precursor mass), our MS-Cluster algorithm[11] scaled this approach for the analysis of tens of millions of spectra and resulted in a significant speed-up of MS^2 database searches (up to 10 fold) while simultaneously increasing the total number of identifications. Soon after, MS-Cluster was extended[44] to be able to process over ≈1.18 billion spectra acquired at Pacific Northwest National Lab over a period of 8 years. This extension served as the foundation for the proposed concept of *spectral archives*,[44] which extend spectral libraries by retaining both identified *and* unidentified spectra in the same way and maintaining information about peptide spectra that are common across species and conditions. Thus archives offer both traditional library spectrum similarity-based search capabilities along with new ways to analyze the data.

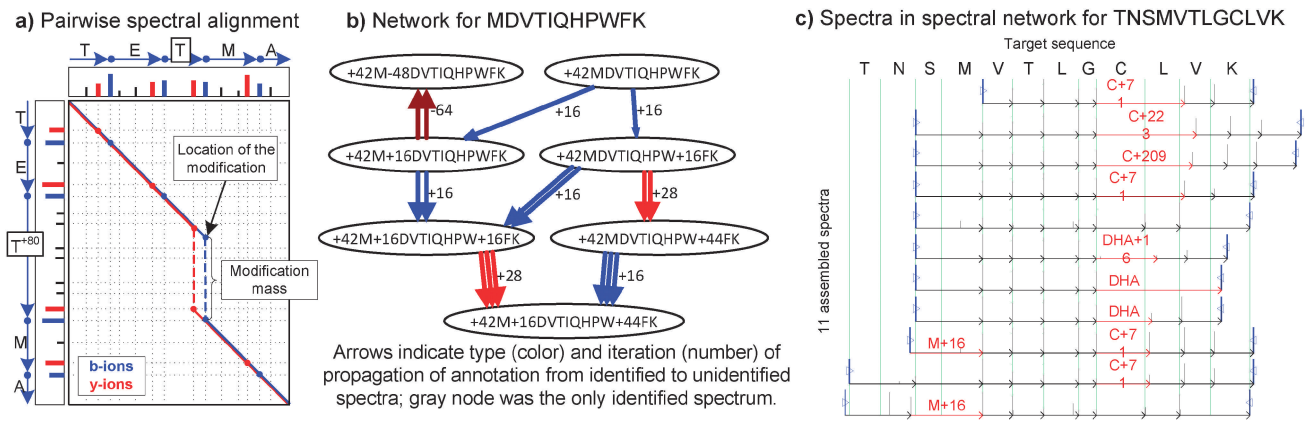


Fig. 1 Discovery and identification of post-translational modifications through spectral networks; (a) Spectral alignment between modified and unmodified variants of the peptide TETMA (*b*-ions shown in blue, *y*-ions in red, blue/red lines track consecutively matched *b*/*y*-ions); (b) Grouped modification states of the peptide MDVTIQHPWFK from a sample of cataractous lenses. Nodes in the spectral network represent individual MS^2 spectra and edges between nodes represent significant spectral alignments such as that shown in part (a); (c) Spectra assembled in the spectral network for TNSMVTLGCLVK with diverse Cysteine modifications on a monoclonal antibody. Each arrow corresponds to the propagation of a sequence and/or PTM from an identified spectrum to an unidentified spectrum (repeated arrows are iterative propagations). Arrow colors correspond to types of modifications transferred.

Spectral networks for analysis of post-translational modifications

Samples of digested proteins often contain multiple overlapping peptides covering the same region of a protein sequence, such as prefix peptides (*e.g.* PEPTI/PEPTIDES), suffix peptides (*e.g.* TIDES/PEPTIDES) or partially-overlapping peptides (*e.g.* PEPTIDES/TIDESHIGH). In addition, most experimental protocols unintentionally generate multiple chemical modifications (*e.g.*, oxidation) and it has been repeatedly shown that existing MS^2 datasets typically contain modified versions for many peptides.[4,45–47] If the peptide sequences were known in advance, determining their overlap would be a straightforward application of the standard sequence alignment algorithms.[48] Conversely, spectral alignment is defined as the alignment of matching peaks between spectra from overlapping peptides.[49,50] This concept is illustrated in Fig. 1a with the matching *b*-ions highlighted in blue. The surprising outcome of spectral alignment, as opposed to sequence alignment, is that even though one does not know the peptide sequences in advance, the sequence information encoded in the masses of the *b*/*y*-ions actually suffices to detect pairs of MS^2 spectra from overlapping peptides. In fact, it turns out that the reliability of spectral alignment allows one to discern the high-scoring true spectral pairs from the many millions of possible spectral pairs in high-throughput proteomics experiments[17,50] Moreover, since each spectrum may align to several other spectra, the set of detected spectral pairs defines a *spectral network* where each node corresponds to a different spectrum and nodes are connected by an edge if the corresponding spectra were found to to be significantly aligned. This concept is illustrated in Fig. 1b–c with spectral networks from human cataractous lens[17] and a monoclonal antibody raised against the B- and T-cell lymphocyte attenuator molecule.[51] Note that since most spectra usually come from non-contiguous protein regions, the consequent outcome of this approach is not a single spectral network but rather multiple spectral networks, one for each set of spectra from overlapping peptides.

In traditional DNA sequence alignment, it often happens that query sequences differ from the reference sequences by the insertion or deletion of one or more nucleotides.[48] While the insertion/deletion of amino acids is also usually allowed when aligning protein sequences, an additional factor needs to be considered when aligning peptides from experimental samples due to the occurrence of post-translational modifications. In fact, multiple groups have shown[16,46,52] that the phenomenon of unexpected modifications is much more widespread than commonly acknowledged. From a sequence alignment perspective, a modification could be modeled by following the modified residue with a special character for each type of modification. Thus, the alignment of a modified peptide PEPT*IDE with its unmodified counterpart PEPTIDE would result in a single difference caused by the insertion of the modification '*' In tandem mass spectrometry, however, a modification of mass *m* conceptually corresponds to the insertion of additional *m* Da in the *b*/*y*-ion series between the ions immediately preceding and following the site of post-translational modification (*i.e.* the mass of the residue becomes larger by mass *m*). Conversely, if the modification causes a loss of *m* Da from the modified residue then the corresponding effect is the subtraction of *m* Da between the ions for the modified residue. When applied to unmodified and modified versions of the same peptide, the role of spectral alignment algorithms[15,17,53] is to (*a*) use the spectrum of the unmodified peptide to determine where to position the modification mass in the spectrum of the modified peptide and (*b*) to assess whether the post-alignment match between the two spectra is significant enough to accept the spectra as a pair of modified/unmodified spectra from the same peptide. Thus, spectral alignment considers every possible spectral pair and every possible location for the mass difference (*i.e.* modification mass) between the aligned spectra. Fig. 1a illustrates the spectral alignment between MS^2 spectra from the peptides TETMA and phosphorylated $TET^{+80}MA$. By requiring a significant match between the aligned spectrum peaks[17] and by placing no restrictions on which modifications to consider, this approach can be used to discover novel or unexpected modifications. In fact, when applied to a set of spectra from cataractous lenses proteins from a 93-year old patient, spectral networks were able to rediscover the modifications identified by database search methods and additionally discovered several novel modification events.[17,46]

When first analyzing a sample possibly containing modified peptides one does not know *a priori* which residues or peptides will be modified. Thus, spectral alignment considers every possible spectral pair and every possible location for the mass difference (*e.g.* modification mass) between the aligned spectra. By requiring a significant match between the aligned spectrum peaks[17] but placing no restrictions on which modifications to consider, this approach can be used to discover novel or unexpected modifications. In fact, when applied to a set of spectra from cataractous lenses proteins from a 93-year old patient, spectral networks were able to rediscover the modifications identified by database search methods and additionally discovered several novel modification events[17,46].

The identification of peptides containing multiple modifications *via* database search is a challenging problem imparted by the combinatorial explosion in the number of possible modification variants for all the peptides in a database.[46,52] Not only can this make the approach much slower, but the increased number of peptide candidates for any given spectrum significantly increases the risk of incorrect identifications. However, samples containing peptides with two or more modifications often also contain variants of the same peptide with only one or no modification. In these cases, we have found that spectral alignment is able to group these related spectra from multiple modification variants of the same peptide into small spectral networks thus increasing confidence in their identity as a related peptide. Fig. 1b illustrates the spectral network for a particular peptide in a sample of cataractous lenses proteins.

By grouping together spectra from multiple variants of the same peptide, spectral networks additionally contribute to the reliable identification of highly modified peptides. While database searching is restricted to matching ion masses between theoretical and observed spectra, spectral networks further capitalizes on the occurrence of common fragment ions at

corresponding masses with similar peak intensities (Fig. 1c). In general, it becomes easier to identify a highly modified peptide if one additionally observes highly-similar spectra from its intermediate modification states. Thus, spectral alignment not only allows one to *discover* unexpected modifications (instead of only identifying expected modifications) but additionally provides an alternative route for identification of highly modified peptides.

4 Shotgun protein sequencing

Current approaches to proteomics focus on the reliable identification of proteins under the assumption that all proteins of interest are known and present in a database. However, the limited availability of sequenced genomes and multiple mechanisms of protein variation often refute this assumption. Well known mechanisms of protein diversity include variable recombination and somatic hypermutation of immunoglobulin genes.[54] The vital importance of some of these novel proteins is directly reflected in the success of monoclonal antibody drugs such as Rituxan, Herceptin and Avastin,[55–57] of which all are derived from proteins that are not directly inscribed in any genome. Similarly, multiple commercial drugs have been developed from proteins obtained from species whose genomes are not known. In particular, peptides and proteins isolated from venom have provided essential clues for drug design[58,59] - examples include drugs for controlling blood coagulation[60–62] and therapeutic treatments for breast[63,64] and ovarian[65] cancer. Despite this vital importance of novel proteins, the mainstream method for protein sequencing is still initiated by restrictive and low-throughput Edman degradation[66,67] - a task made difficult by protein purification procedures, post-translational modifications and blocked protein N-termini. These problems gain additional relevance when one considers the unusually high level of variability and post-translational modifications in venom proteins.[68,69]

Conceptually, sequencing a protein from a set of MS^2 spectra can be described by a simple analogy. Imagine a jewellery box with many identical copies of a specific model of bead necklaces. Although all the beads are identical, this model is characterized by having irregular distances between consecutive beads–the set of inter-bead distances is initially chosen by the designer and all necklaces are then made using exactly the same specification. Now assume that one day you open your jewellery box and realize that someone has vandalized all the necklaces by cutting them to fragments at randomly chosen bead positions. Can you recover the original design of this model of necklaces, as specified by the set of consecutive inter-bead distances? In this allegory inter-bead distances correspond to amino acid masses and beads correspond to MS^2 fragmentation points (between consecutive amino acids). MS^2 data add more than a few difficulties to this necklace assembly problem; for example, most peaks in MS^2 spectra do not correspond to any fragment ions (extra beads) and many fragment ions do not result in any peaks (missing beads). Nevertheless, Fig. 2 presents an example of assembled MS^2 spectra resulting in a 22 amino acid long segment of a monoclonal antibody.[51]

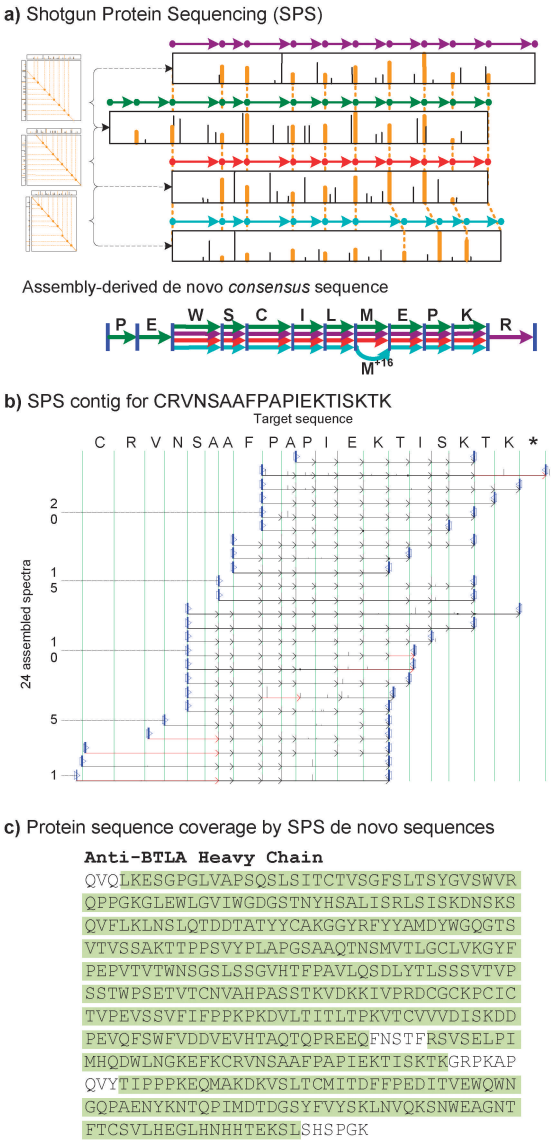


Fig. 2 Shotgun Protein Sequencing (SPS) *via* assembly of tandem mass spectra; (a) Spectral alignment between spectra for peptide WSCILMEPKR (purple), PEWSCILMEPKR (green), WSCILMEPK (red), WSCILMoxEPK (cyan); Mox represents oxidized Methionine. Matching peaks in spectral alignments become pairwise gluing instructions between every pair of aligned spectra. (b) Protein contig resulting from 24 spectra from a monoclonal antibody (aBTLA heavy chain). Each spectrum is shown superimposed with a sequence of arrows indicating its sequence of recovered masses; modified variants of the consensus sequence are indicated by red arrows (6 different modifications on 7 spectra). (c) The complete aBTLA heavy chain sequence recovered by Comparative SPS;[57] highlighted sections were covered by protein contigs (95% coverage) and the missing amino acids were obtained from homologous protein sequences.

Shotgun Protein Sequencing (*SPS*) is a *de novo* sequencing approach[15] that utilizes multiple MS^2 spectra from overlapping peptides generated using non-specific proteases or multiple proteases with different specificities.[70–74] The original approach was based on the overlap → layout → consensus approach to assembly and shown to be efficient for the assembly of a single purified unmodified protein. However,

practical applications (like sequencing snake venoms) require applicability to mixtures of modified proteins. In fact, most MS^2 samples contain both modified and unmodified versions for many peptides, including biological and chemical modifications both native and introduced during sample preparation. Sequence variations and post-translational modifications present a formidable algorithmic challenge for assembly algorithms as the performance of the original SPS approach[15] steeply degraded as soon as even a small percentage of the spectra are from modified peptides. To use the beads analogy, the necklace puzzle becomes very difficult if in addition to the canonical necklaces (non-modified proteins), the jewellery box also contains some necklaces that deviate from the designer's specification (modified proteins). Building on spectral networks algorithms for analysis of post-translational modifications based on alignment of spectra from modified and unmodified peptide variants,[17,50] we showed how to integrate these alignments into Shotgun Protein Sequencing to derive a completely new form of spectral assembly. This utilized a generalized notion of *ABruijn graphs* (originally proposed in the context of DNA fragment assembly[75]) for the assembly of MS^2 spectra from overlapping, modified and unmodified peptides into *contigs* (sets of aligned spectra from overlapping peptides, see Fig. 2), where each contig then capitalizes on the corroborating evidence from the assembled spectra to yield a high-quality *consensus de novo* sequence. As a result, SPS consensus *de novo* sequences were found to be twice as accurate as sequences derived from single spectra (1 mistake per 10 *vs.* 5 amino acid predictions) while yielding sequences that were much longer that single-peptide/spectrum could support (up to 24 AA long).

Recently this paradigm was extended in two distinct directions. First, we capitalized on homology between SPS long/accurate *de novo* sequences and known sequences to deliver the first automated full-length protein sequencing approach (Comparative SPS[57]) and demonstrated it with database-assisted *de novo* sequencing of two monoclonal antibodies. Spectral networks also underlie the related work of Castellana *et al.*,[76,77] who proposed an effective method for sequencing monoclonal antibodies with database-guided iterative alignment + assembly of spectra from overlapping peptides. Both of these methods rely upon the existence of a homologous database. To reduce this dependence, we have since developed MetaSPS[78] algorithms for assembling SPS contigs into *meta-contigs* (sets of overlapping contigs). These methods now deliver *de novo* sequences over 100 AA long at sequencing error rates as low as 1 mistake per 50 predicted amino acids without requiring homology to known sequences, which demonstrates the feasibility of fully-automated *de novo* protein sequencing with unidentified MS^2 spectra. It is expected that the performance of these algorithms will only improve as new types of mass spectrometry data (*e.g.*, Electron Transfer Dissociation) are also incorporated in SPS and spectral networks approaches.

5 Spectral networks for non-ribosomal, cyclic peptides

The central dogma of biology (translation of template mRNA into proteins/peptides) is not the only mechanism for cells to assemble amino acids into peptides. The alternative *Non Ribosomal Peptide Synthesis* is performed by a large multi-enzyme complex (called Non Ribosomal Peptide Synthetase or NRPS) that represents both the biosynthetic machinery and the mRNA-free template for the biosynthesis of secondary metabolites.[79–81] NRPS gene clusters produce relatively short (up to 50 AA) nonribosomal peptides (NRP) that are not directly inscribed in the genomic DNA and thus cannot be inferred with traditional DNA-based sequencing techniques. NRPs are of tremendous pharmacological importance since they were optimized during millions of years of evolution to play important roles in chemical defense and communication for producing organisms. Starting from penicillin, NRPs and other natural products (*i.e.* secondary metabolites) have an unparalleled track record in pharmacology: 9 out of the top 20 best-selling drugs were either inspired by or derived from natural products. NRPs have some naturally evolved features that are applicable to the modulation of protein function in human systems, making them excellent lead compounds for the development of novel pharmaceutical agents. In particular, NRPs include antibiotics (penicillin, cephalosporin, vancomycin, *etc.*), immunosuppressors (cyclosporine, tacrolimus, sirolimus), antiviral agents (luzopeptin A), antitumor agents (bleomycin), toxins (thaxtomin), and many peptides with yet unknown functions.

When DNA sequencing is not available, biologists use either Edman degradation or tandem mass spectrometry (MS^2) to sequence ribosomal peptides. However, neither of these approaches works for nonribosomal peptides since they differ from ribosomal peptides in many respects: (a) they often represent non-linear structures of amino acids (*e.g.*, cyclic, tree-like, and branch-cyclic peptides), (b) they often contain non-standard amino acids increasing the number of possible building blocks from 20 to several hundred, (c) they often have a non-standard backbone, and (d) they are often modified. Each of these complications renders traditional Edman degradation and MS^2 peptide sequencing approaches useless, leaving NMR as the only technology capable of analyzing NRPs.[82–85] The use of NMR for NRP sequencing is time-consuming, difficult to automate (there are currently no software tools for automatic interpretation of NRPs from NMR data), and error-prone (see[85,86] for examples of errors in NMR sequencing). In addition, the abundance of these specialized compounds *in vivo* is often very low requiring extensive raw biological material in order to purify enough of the compound to perform 2D NMR for structure elucidation. As a result, the extremely difficult process of total chemical synthesis remained one of the only reliable way to sequence and validate NRPs.[87]

Having shown how multi-stage mass spectrometry (MS^n) can improve *de novo* sequencing accuracy for linear peptides,[88] we then extended spectral networks algorithms using a combination of experimental and computational protocols to enable a mass-spectrometry based approach for *de novo* sequencing of cyclic peptides.[89] The NRP-Sequencing algorithm discovers

amino acid masses and reconstructs cyclic peptide sequences directly from a single MS^3 spectrum and MS^4 MS^{-5} spectra are used to rescore all putative MS^3 reconstructions. The NRP-Assembly approach assembles MS^4 MS^{-5} spectra, similarly to what was described above for Shotgun Protein Sequencing, and further integrates the resulting contig with the MS^3 spectrum and all non-assembled spectra. These algorithmic foundations were further extended as more data became available[90] and we were able to show how these tools can conserve significant efforts using several marine cyanobacterial cyclic peptides. In particular, Cyanopeptide X was an unknown bioactive molecular whose identity was elucidated using the very time intensive workflow of isolating, purifying and collecting 2D NMR data to obtain the structure.[91] However, using very small amounts of raw material, sequencing of MS^2 data using our cyclic peptide annotation algorithms revealed that this compound was related to dolastatin 11 (reversed amino acid sequence with a single modification) and majusculamide C with identical scores, which provided great insight into the nature of the structure with very little time investment. The compound turned out to be desmethoxymajusculamide C and a full report on its structure as determined by NMR is now available.[92] Another example was compound 879, which was initially assumed to be a novel compound but was later found to be already known during the patent application. Our analysis could dereplicate the spectrum of compound 879 as the known NRP neoviridogrisen and could thus have saved the three years of effort it took to determine the structure.

6 Spectral networks for any type of molecules

Microbes use secreted factors to interact, communicate and manipulate their local environment and neighboring cell populations in a process known as metabolic exchange. By employing a wide breadth of molecules ranging from signaling compounds to defensive metabolites, metabolic exchange dictates not only basic microbial behavior such as biofilm formation, sporulation and motility, but also social interactions such as syntrophy and quorum sensing which enables microbes to establish communities.[93–100] Despite these secreted factors having a major impact on the phenotypic development of microbial populations, there is a lack of tools that enable scientists to probe the chemistry of microbial colonies in a direct manner, let alone of live microbial colonies. Currently the chemistry of microbes is studied indirectly and, in general, on single molecules—an effort with a significant time and monetary investment. Furthermore, since organisms are not static entities, it is important to be able to monitor chemical exchanges temporally and spatially as both the timing of production and the distribution of metabolic exchange factors within microbial populations can provide valuable insight into the function of these molecules.

As with peptide-based spectral networks, molecular spectral networks[101] start with raw MS^2 data acquired from one or more microbial species, irrespective of the number of spectra or mass spectrometry runs. Then, similarly to the algorithm illustrated in Fig. 1a), pairs of MS^2 spectra from related molecules are detected using *structure-independent* spectral alignment to find spectra with significantly-similar fragmentation patterns, regardless of whether the spectra are identified in advance or not. By avoiding peptide-specific fragmentation models and assumptions, structure-independent spectral alignment reveals molecular networks containing not only spectra of peptides but also primary and secondary metabolites, non-linear natural-products, lipids, glycans, and other classes of molecules. Fig. 4 shows a molecular spectral network for *Bacillus subtilis* 3610 and the chemical structures for several compounds corresponding to specific highlighted subcomponents of the whole network.

7 Conclusions

The spectral networks paradigm is founded on two core principles beyond mainstream approaches: (1) it is more efficient to match unidentified spectra to reference or other unidentified spectra than to reference sequences and (2) consensus interpretation of sets of related spectra is more reliable than identification of one spectrum at a time. In both instances it is relatively easy to see how these principles go beyond the potential of mainstream approaches. Reference spectra are also associated with reference sequences and annotations so additional knowledge of previously observed MS^2 fragmentation patterns can only improve identification algorithms. In addition to improving traditional peptide identification approaches, we have also shown how spectrum matching and library search algorithms[40] can support new directions such as identification of mixed spectra with more than one peptide beyond the state of the art in comparable database search approaches.[40,102] Also, having sets of spectra from related versions of the same compound (*e.g.*, modified/unmodified, N C^{-1}-term extensions, CID/ETD/HCD/MS^n spectra, *etc.*) significantly increases signal-to-noise ratios by providing more signal/fragmentation ions and averaging out inconsistent noise across all related spectra. As opposed to other multi-spectrum peptide identification and *de novo* sequencing approaches, spectral networks algorithms eliminate the need for the sets of spectra to be determined by the mass spectrometry instrument (*e.g.*, as in MS^2 MS^{-3} protocols) and allow for correlation, alignment and assembly of spectra across multiple peptide sequence variations, post-translational modification, experimental conditions and even multiple species. The *de novo* sequencing subset of spectral networks algorithms (Shotgun Protein Sequencing, or SPS) clearly illustrates the potential of this approach by uniquely delivering *de novo* sequences longer than single-spectrum peptide sequences[103] which now span over 100 amino acids at an average accuracy of less than one sequencing error per 50 AA.[78] When combined with error-tolerant database search algorithms, SPS also enabled the first automated full-length protein sequencing approach, as demonstrated by our *de novo* sequencing of multiple monoclonal antibodies directly from a protein extract.[57]

As with spectral matching and library search, the potential of spectral networks algorithms extends beyond the scope of significantly improving on traditional uses of mass spectrometry data. Despite the significant clinical importance of natural products drug discovery, automated analysis of their

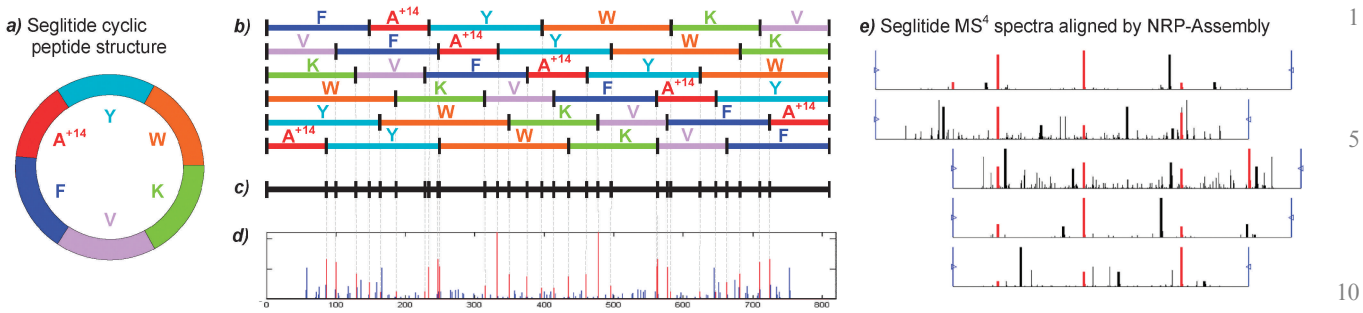


Fig. 3 Analysis of the cyclic peptide Seglitide. (a) The circular structure of Seglitide is schematically illustrated with each residue represented by a different color (slice sizes not scaled to corresponding masses of the residues). A^{+14} denotes a non-standard residue with integer mass 71 + 14 = 85 Da. (b) MS^2 fragmentation of Seglitide generates up to 6 linear peptides representing different rotated variants of the same cyclic peptide. (c) Theoretical spectrum for Seglitide by superposition of the fragment masses of the linearized peptides. (d) Experimental spectrum of Seglitide resulting from a mixture of 6 linear peptides (the peaks corresponding to fragment ions are shown in red). (e) Spectral network from assembled Seglitide MS^n spectra and used for *de novo* sequencing with unknown amino acid masses.

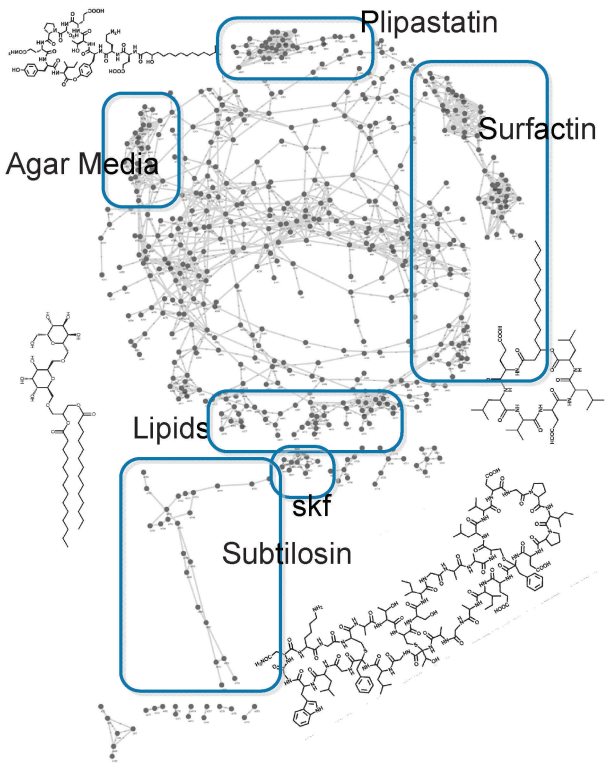


Fig. 4 Molecular spectral network of a partial Bacillus subtilis secretome; nodes indicate MS^2 spectra of initially-unknown compounds of any class of molecules (no peptide-specific assumptions were made), and edges indicate significant similarity between the MS^2 fragmentation patterns of different spectra, mostly between intermediates/variants of the same compounds. Selected molecular structures are shown in black overlaid with the network and next to the correspondingly highlighted network clusters.

mass spectra has always been tremendously challenging since these are often non-ribosomal and have no genomic propeptide template, are assembled with non-standard and heavily-modified amino acids and almost always have non-linear structures such as multi-cyclic, branched-cyclic and others–each of which renders traditional database search and *de novo* sequencing algorithms essentially useless. Using a combination of new mass spectrometry protocols and novel spectral networks algorithms, we showed[89] how amino acid masses can be discovered directly from the data and how spectra of cyclic peptides can be assembled into accurate *de novo* sequences, a direction that was later explored for the analysis of several more novel natural products.[90,91] Building on these results and recent advances,[101] the scope of spectral networks analysis has now been extended to the analysis of tandem mass spectra for any type of molecules by aligning spectrum fragmentation patterns without any prior assumptions on molecular structure or composition. As such, preliminary results indicate that the spectral networks paradigm may serve as the foundation to organize and search a mass spectrometry-centric view of the complete biomolecular space.

Being a relatively new paradigm,[15,17,103] the field of spectral networks analysis of tandem mass spectrometry data remains rich with open computational problems that stand to substantially benefit from additional developments in spectral matching, alignment, assembly and consensus interpretation. These and related developments continue to be proposed in closely related fields[29,40,104] and are expected to have a substantial impact on the quality and extent of future spectral networks repositories and tools.

Acknowledgements

The Center for Computational Mass Spectrometry at UCSD is supported by the National Institutes of Health Grant 1-P41-RR024851 from the National Center for Research Resources. The P.C.D. laboratory is supported by US National Institutes of Health grants GM097509, GM094802, GM086283 and AI095125 and the Keck foundation.

References

1 J. Eng, M. AL and J. Yates, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 976–989.
2 D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, *Electrophoresis*, 1999, **20**, 3551–3567.
3 R. Craig and R. C. Beavis, *Bioinformatics*, 2004, **20**, 1466–1467.
4 S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. Pevzner and V. Bafna, *Anal. Chem.*, 2005, **77**, 4626–4639.

5 B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, *Rapid Commun. Mass Spectrom.*, 2003, **17**, 2337–234.
6 A. M. Frank and P. A. Pevzner, *Anal. Chem.*, 2005, **77**, 964–973.
7 B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem and J. M. Buhmann, *Anal. Chem.*, 2005, **77**, 7265–7273.
8 L. Mo, D. Dutta, Y. Wan and T. Chen, *Anal. Chem.*, 2007, **79**, 4870–4878.
9 A. Keller, A. Nesvizhskii, E. Kolker and R. Aebersold, *Anal. Chem.*, 2002, **74**, 5383–5392.
10 J. E. Elias and S. P. Gygi, *Nat. Methods*, 2007, **4**, 207–214.
11 A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith and P. A. Pevzner, *J. Proteome Res.*, 2008, **7**, 113–122.
12 R. Craig, J. C. Cortens, D. Fenyo and R. C. Beavis, *J. Proteome Res.*, 2006, **5**, 1843–1849.
13 B. E. Frewen, G. E. Merrihew, C. C. Wu, W. S. Noble and M. J. MacCoss, *Anal. Chem.*, 2006, **78**, 5678–5684.
14 H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold, *Proteomics*, 2007, **7**, 655–667.
15 N. Bandeira, H. Tang, V. Bafna and P. Pevzner, *Anal. Chem.*, 2004, **76**, 7221–7233.
16 M. M. Savitski, M. L. Nielsen and R. A. Zubarev, *Mol. Cell. Proteomics*, 2006, **5**, 935–948.
17 N. Bandeira, D. Tsur, A. Frank and P. Pevzner, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 6140–6145.
18 C. R. Robertson, S. P. Flynn, H. S. White and G. Bulaj, *Nat. Prod. Rep.*, 2011, **28**, 741–762.
19 G. F. King, *Expert Opin. Biol. Ther.*, 2011, **11**, 1469–1484.
20 M. L. Colgrave, A. G. Poth, Q. Kaas and D. J. Craik, *Biopolymers*, 2010, **94**, 592–601.
21 A. S. Jaggi and N. Singh, *CNS Neurol Disord Drug Targets*, 2011, **10**, 589–609.
22 F. R. Depontieu, J. Qian, A. L. Zarling, T. L. McMiller, T. M. Salay, A. Norris, A. M. English, J. Shabanowitz, V. H. Engelhard, D. F. Hunt and S. L. Topalian, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 12073–12078.
23 F. Klug, M. Miller, H. H. Schmidt and S. Stevanovi, *Curr. Pharm. Des.*, 2009, **15**, 3221–3236.
24 J. Zheng, R. J. Sugrue and K. Tang, *Anal. Chim. Acta*, 2011, **702**, 149–159.
25 K. J. Pflughoeft and J. Versalovic, *Annu. Rev. Pathol.*, 2011.
26 H. Lam, E. W. Deutsch and R. Aebersold, *J. Proteome Res.*, 2010, **9**, 605–610.
27 A. I. Nesvizhskii, *J. Proteomics*, 2010, **73**, 2092–2123.
28 C. Y. Yen, K. Meyer-Arendt, B. Eichelberger, S. Sun, S. Houel, W. M. Old, R. Knight, N. G. Ahn, L. E. Hunter and K. A. Resing, *Mol. Cell. Proteomics*, 2009, **8**, 857–869.
29 C. Y. Yen, S. Houel, N. G. Ahn and W. M. Old, *Mol. Cell. Proteomics*, 2011, **10**, in press.
30 Y. Kim, S. Bark, V. Hook and N. Bandeira, *Bioinformatics*, 2011, **27**, 2772–2773.
31 V. Hook, S. Bark, N. Gupta, M. Lortie, W. D. Lu, N. Bandeira, L. Funkelstein, J. Wegrzyn, D. T. O'Connor and P. Pevzner, *AAPS J.*, 2010, **12**, 635–645.
32 A. Bora, S. P. Annangudi, L. J. Millet, S. S. Rubakhin, A. J. Forbes, K. N. L. , M. U. Gillette and J. V. Sweedler, *J. Proteome Res.*, 2008, **7**, 4992–5003.
33 L. D. Fricker, *Endocrinology*, 2007, **148**, 4185–4190.
34 L. Li and J. V. Sweedler, *Annu. Rev. Anal. Chem.*, 2008, **1**, 451–483.
35 S. M. K. Skold, A. Nilsson, M. Falth, P. Svenningsson and P. E. A. Andren, *Biochem. Soc. Trans.*, 2007, **35**, 588–593.
36 D. Phanstiel, J. Brumbaugh, W. T. Berggren, K. Conard, X. Feng, M. E. Levenstein, G. C. McAlister, J. A. Thomson and J. J. Coon, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 4093–4098.
37 J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin and J. R. Yates, *Nat. Methods*, 2004, **1**, 39–45.
38 C. Masselon, L. Pasa-Tolic', S. W. Lee, L. Li, G. A. Anderson, R. Harkewicz and R. D. Smith, *Proteomics*, 2003, **3**, 1279–1286.
39 A. B. Chakraborty, S. J. Berger and J. C. Gebler, *Rapid Commun. Mass Spectrom.*, 2007, **21**, 730–744.
40 J. Wang, J. Pérez-Santiago, J. E. Katz, P. Mallick and N. Bandeira, *Mol. Cell. Proteomics*, 2010, **9**, 1476–1485.
41 D. Tabb, M. MacCoss, C. Wu, S. Anderson and J. r. Yates, *Anal. Chem.*, 2003, **75**, 2470–2477.
42 I. Beer, E. Barnea, T. Ziv and A. Admon, *Proteomics*, 2004, **4**, 950–960.
43 D. L. Tabb, M. Thompson, G. Khalsa-Moyers, N. VerBerkmoes and W. McDonald, *J. Am. Soc. Mass Spectrom.*, 2005, **16**, 1250–1261.
44 A. M. Frank, M. E. Monroe, A. R. Shah, J. J. Carver, N. Bandeira, R. J. Moore, G. A. Anderson, R. D. Smith and P. A. Pevzner, *Nat. Methods*, 2011, **8**, 587–591.
45 E. Hunyadi-Gulyas and K. Medzihradszky, *Drug Discovey Today: Targets–mass spectrometry in proteomics supplement*, 2004, **3**, 3–10.
46 D. Tsur, S. Tanner, E. Zandi, V. Bafna and P. A. Pevzner, *Nat. Biotechnol.*, 2005, **23**, 1562–1567.
47 P. A. Wilmarth, S. Tanner, S. Dasari, S. R. Nagalla, M. A. Riviere, V. Bafna, P. A. Pevzner and L. L. David, *J. Proteome Res.*, 2006, **5**, 2554–2566.
48 T. F. Smith and M. S. Waterman, *J. Mol. Biol.*, 1981, 195–197.
49 P. Pevzner, V. Dancík and C. Tang, *J. Comput. Biol.*, 2000, **7**, 777–787.
50 N. Bandeira, D. Tsur, A. Frank and P. Pevzner, *Proceeding of the Tenth Annual International Conference in Research in Computational Molecular Biology (RECOMB 2006)*, 2006, pp. 363–378.
51 N. Bandeira, V. Pham, P. Pevzner, A. D. and L. J. R. , *Nat. Biotechnol.*, 2008, **26**, 1336–8.
52 N. S. , B. N and P. E. , *Mol. Cell. Proteomics*, 2011, in press.
53 P. Pevzner, Z. Mulyukov, V. Dancik and C. Tang, *Genome Res.*, 2001, **11**, 290–299.
54 P. J. Gearhart, *Nature*, 2002, **419**, 29–31.
55 M. Wiles and P. Andreassen, *Drug Discov World*, 2006, **Fall 2006**, 17–23.
56 J. S. Haurum, *Drug Discovery Today*, 2006, **11**, 655–660.
57 N. Bandeira, V. Pham, P. Pevzner, D. Arnott and J. R. Lill, *Nat. Biotechnol.*, 2008, **26**, 1336–1338.
58 R. J. Lewis and M. L. Garcia, *Nat. Rev. Drug Discovery*, 2003, **2**, 790–802.
59 A. M. Pimenta and M. E. De Lima, *J. Pept. Sci.*, 2005, **11**, 670–676.
60 J. Joseph and R. Kini, *Curr. Drug Targets: Cardiovasc. & Haematol. Disord.*, 2004, **4**, 397–416.
61 S. Swenson, C. Toombs, L. Pena, J. Johansson and F. Markland, *Curr. Drug Targets: Cardiovasc. & Haematol. Disord.*, 2004, **4**, 417–435.
62 R. Kini, V. Rao and J. Joseph, *Haemostasis*, 2001, **31**, 218–224.
63 S. Swenson, F. Costa, R. Minea, R. Sherwin, W. Ernst, G. Fujii, D. Yang and F. Markland, *Mol. Cancer. Ther.*, 2004, **3**, 499–511.
64 S. K. Pal, A. Gomes, S. C. Dasgupta and A. Gomes, *Indian. J. Exp. Biol.*, 2002, **40**, 1353–1358.
65 F. Markland, K. Shieh, Q. Zhou, V. Golubkov, R. Sherwin, V. Richters and R. Sposto, *Haemostasis*, 2001, **31**, 183–191.
66 A. Zugasti-Cruz, M. Maillo, E. López-Vera, A. Falcón, E. P. Heimer de la Cotera, B. M. Olivera and M. B. Aguilar, *Peptides*, 2006, **27**, 506–511.
67 Y. Ogawa, R. Yanoshita, U. Kuch, Y. Samejima and D. Mebs, *Toxicon*, 2004, **43**, 855–858.
68 O. Buczek, G. Bulaj and B. M. Olivera, *Cell. Mol. Life Sci.*, 2005, **62**, 3067–3079.
69 A. M. Pimenta, B. Rates, C. Bloch, P. C. Gomes, M. M. Santoro, M. E. de Lima, M. Richardson and M. d. o. N. Cordeiro, *Rapid Commun. Mass Spectrom.*, 2005, **19**, 31–37.
70 R. Johnson and K. Biemann, *Biochemistry*, 1987, **26**, 1209–1214.
71 A. A. Klammer and M. J. MacCoss, *J. Proteome Res.*, 2006, **5**, 695–700.
72 J. Englander, C. Del Mar, W. Li, S. Englander, J. Kim, D. Stranz, Y. Hamuro and V. Woods, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 7057–7062.
73 M. MacCoss, W. McDonald, A. Saraf, R. Sadygov, J. Clark, J. Tasto, K. Gould, D. Wolters, M. Washburn, A. Weiss, J. Clark and J. Yates, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 7900–7905.
74 V. Pham, W. J. Henzel, D. Arnott, S. Hymowitz, W. N. Sandoval, B. T. Truong, H. Lowman and J. R. Lill, *Anal. Biochem.*, 2006, **352**, 77–86.

75 P. A. Pevzner, H. Tang and G. Tesler, *Genome Res*., 2004, **14**, 1786–1796.
76 N. E. Castellana, V. Pham, D. Arnott, J. R. Lill and V. Bafna, *Mol. Cell. Proteomics*, 2010, **9**, 1260–1270.
77 N. E. Castellana, K. McCutcheon, V. C. Pham, K. Harden, A. Nguyen, J. Young, C. Adams, K. Schroeder, D. Arnott, V. Bafna, J. L. Grogan and J. R. Lill, *Proteomics*, 2011, **11**, 395–405.
78 A. Guthals, K. Clauser and N. Bandeira, *(submitted)*, 2012.
79 S. A. Sieber and M. A. Marahiel, *Chem. Rev*., 2005, **105**, 715–738.
80 P. C. Dorrestein and N. L. Kelleher, *Nat. Prod. Rep*., 2006, **23**, 893–918.
81 M. Welker and H. Von Doehren, *FEMS Microbiol. Rev*., 2006, **30**, 530–563.
82 B. G. Butcher and J. D. Helmann, *Mol. Microbiol*., 2006, **60**, 765–782.
83 D. E. Williams, P. Austin, A. R. Diaz-Marrero, R. V. Soest, T. Matainaho, C. D. Roskelley, M. Roberge and R. J. Andersen, *Org. Lett*., 2005, **7**, 4173–4176.
84 H. Luesch, P. G. Williams, W. Y. Yoshida, R. E. Moore and V. J. Paul, *J. Nat. Prod*., 2002, **65**, 996–1000.
85 T. Hamada, S. Matsunaga, G. Yano and N. Fusetani, *J. Am. Chem. Soc*., 2005, **127**, 110–118.
86 C. M. Ireland, A. R. Durso, R. A. Newman and M. P. Hacker, *J. Org. Chem*., 1982, **47**, 360–361.
87 K. Kurosawa, K. Matsuura and N. Chida, *Tetrahedron Lett*., 2005, **46**, 389–392.
88 N. Bandeira, J. Olsen, M. Mann and P. Pevzner, *Bioinformatics (ISMB 2008 special issue)*, 2008, **24**, i416–i423.
89 N. Bandeira, J. Ng, D. Meluzzi, R. Linington, P. Dorrestein and P. Pevzner, *Proceedings of the Twelfth Annual International Conference in Research in Computational Molecular Biology (RECOMB 2008)*, 2008, pp. 181–195.
90 J. Ng, N. Bandeira, W. Liu, R. Linington, P. Dorrestein and P. Pevzner, *Nat. Methods*, 2009, **6**, 596–599.
91 W. Liu, J. Ng, D. Meluzzi, N. Bandeira, M. Gutierrez, T. Simmons, A. Schultz, R. Linington, B. Moore, W. Gerwick, P. Pevzner and P. Dorrestein, *Anal. Chem*., 2009, **81**, 4200–4209.
92 T. L. Simmons, L. M. Nogle, J. Media, F. A. Valeriote, S. L. Mooberry and W. H. Gerwick, *J. Nat. Prod*., 2009, **72**, 1011–1016.
93 V. V. Phelan, W.-T. Liu, K. Pogliano and P. C. Dorrestein, *Nat. Chem. Biol*., 2011, **8**, 26–35.
94 W. L. Ng and B. L. Bassler, *Annu. Rev. Genet*., 2009, **43**, 197–222.
95 P. D. Straight and R. Kolter, *Annu. Rev. Microbiol*., 2009, **63**, 99–118.
96 A. E. Little, C. J. Robinson, S. B. Peterson, K. F. Raffa and J. Handelsman, *Annu. Rev. Microbiol*., 2008, **62**, 375–401.
97 D. Lpez and R. Kolter, *FEMS Microbiol. Rev*., 2010, **34**, 134–149.
98 G. Yim, H. H. Wang and J. Davies, *Phil. Trans. R. Soc. Lond*., 2007, **B362**, 1195–1200.
99 E. A. Shank and R. Kolter, *Curr. Opin. Microbiol*., 2011, **14**, 741–747.
100 D. Romero, M. F. Traxler, D. Lpez and R. Kolter, *Chem. Rev*., 2011, **111**, 5492–5505.
101 J. Watrous, P. J. Roachc, T. Alexandrov, B. S. Heathc, J. Y. Yang, R. Kersten, M. Voort, K. Pogliano, H. Gross, J. Raaijmakersi, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *(submitted)*, 2012.
102 J. Wang, P. E. Bourne and N. Bandeira, *Mol. Cell. Proteomics*., 2011, in press.
103 N. Bandeira, K. Clauser and P. Pevzner, *Mol. Cell. Proteomics*, 2007, **6**, 1123–34.
104 H. Lam, *Mol. Cell Proteomics*, 2011, **10**, in press.