J Ind Microbiol Biotechnol
DOI 10.1007/s10295-015-1675-9

BIOTECHNOLOGY METHODS

Informatic search strategies to discover analogues and variants of natural product archetypes

Chad W. Johnston[1,2] · Alex D. Connaty[1,2] · Michael A. Skinnider[1,2] · Yong Li[1,2] · Alyssa Grunwald[3,4] · Morgan A. Wyatt[1,2] · Russell G. Kerr[3,4] · Nathan A. Magarvey[1,2]

Received: 17 June 2015 / Accepted: 13 August 2015

Abstract Natural products are a crucial source of antimicrobial agents, but reliance on low-resolution bioactivity-guided approaches has led to diminishing interest in discovery programmes. Here, we demonstrate that two in-house automated informatic platforms can be used to target classes of biologically active natural products, specifically, peptaibols. We demonstrate that mass spectrometry-based informatic approaches can be used to detect natural products with high sensitivity, identifying desired agents present in complex microbial extracts. Using our specialised software packages, we could elaborate specific branches of chemical space, uncovering new variants of trichopolyn and demonstrating a way forward in mining natural products as a valuable source of potential pharmaceutical agents.

Keywords Natural products · Peptaibols · LC–MS/MS · Nonribosomal peptides · Informatic platforms

Special Issue: Natural Product Discovery and Development in the Genomic Era. Dedicated to Professor Satoshi Ōmura for his numerous contributions to the field of natural products.

Electronic supplementary material The online version of this article (doi:10.1007/s10295-015-1675-9) contains supplementary material, which is available to authorized users.

✉ Nathan A. Magarvey
magarv@mcmaster.ca

[1] Department of Biochemistry and Biomedical Sciences, The Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON L8N 3Z5, Canada

[2] Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON L8N 3Z5, Canada

[3] Department of Chemistry, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI C1A 4P3, Canada

[4] Department of Biomedical Sciences, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI C1A 4P3, Canada

Introduction

Microbial natural products have been the most important source of antimicrobial agents for nearly a century, providing the basis for ~80 % of the antibiotics used today [18]. The golden age of natural product discovery (occurring roughly between the 1950s and 1960s) provided thousands of valuable small molecules that could be tracked and isolated based on their overt biological or antimicrobial activity [10]. However, over reliance on bioactivity-guided approaches to antimicrobial identification has resulted in a high rate of rediscovery of common, abundant compounds, fostering disinterest in antimicrobial development and leading to the abandonment of industrial natural products discovery [9, 10]. In light of the increasing occurrence of antimicrobial resistance, natural products are regaining attention as a source of novel drugs, although new techniques will be required to modernise the methods of their discovery [11].

Previously, we described the development of an informatic algorithm for identifying known peptide natural products from LC–MS data of complex microbial culture extracts, known as iSNAP [7]. This algorithm used a series of statistical scoring functions to match observed MS/MS fragmentation spectra to hypothetical MS/MS spectra of known compounds that had been fragmented in silico, and was shown to be a reliable means of identifying known peptide natural products with an exceptionally low rate of false

Published online: 08 September 2015

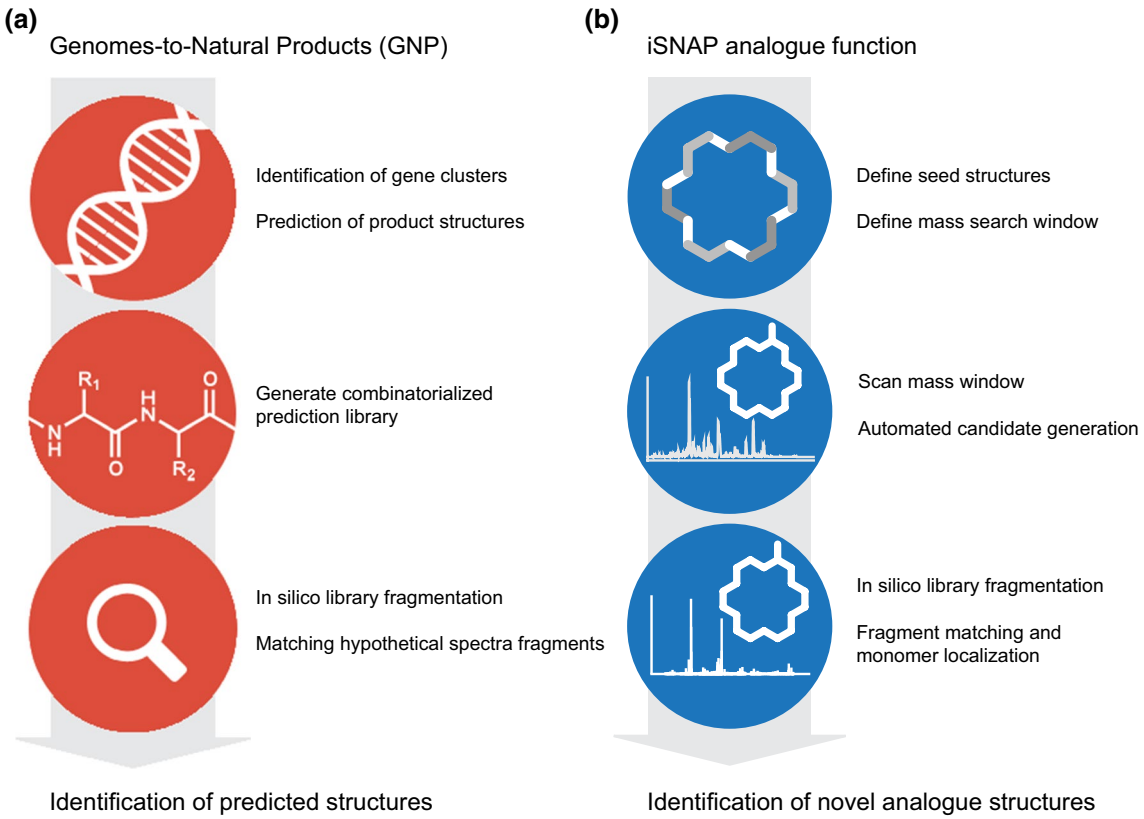

Fig. 1 Automated workflows for GNP and the iSNAP analogue function. **a** The automated GNP workflow can be initiated at any stage, but typically begins with the automated identification of NRP and PK biosynthetic gene clusters, generating predictions of their products. Predictions are combinatorialized in an online user-interface, and the corresponding library is fragmented in silico and used to search real MS/MS data for related structures. **b** The iSNAP analogue function begins with defining desired seed structures for which analogues are identified from MS data, searching within a defined mass window. Candidate structures are generated for hits within this window, and MS/MS fragment matching is used to assess statistical validity of hits and localise structural changes

positive discovery. Recently, we detailed two advanced versions of this algorithm that could detect new natural products as well as known ones (Fig. 1). The Genomes-to-Natural Products platform (GNP) is a fully automated web application capable of detecting, predicting, combinatorializing, and identifying nonribosomal peptide (NRP) and polyketide (PK) products based on genetic information and LC–MS data (Johnston et al. Accepted). This approach matches hypothetical spectral fragments from libraries of combinatorialized structures to real MS/MS data and identifies similar structures, facilitating the identification of molecules based on genetic predictions or inferred biosynthetic variability (Fig. 1a). Meanwhile, the iSNAP analogue function is an addition to the original algorithm that allows it to identify the locations and structures of analogues of dereplicated natural products, revealing site-specific modifications and detailing families of related structures with high accuracy (Yang et al. resubmitted with revisions). This second approach searches a user-defined mass window around a designated or dereplicated seed structure, using MS/MS data to localise the mass gap between the seed and analogue to a specific monomer (Fig. 1b), thus detailing novel congeners in an automated fashion. Having detailed these two advanced informatic software packages, we now seek to apply these tools to identify desirable antimicrobial natural products for treating high priority diseases, such as systemic fungal infections.

Beyond the four clinically used antifungal drug classes (allylamines, azoles, polyenes, and echinocandins [15]), there remain a number of microbial natural products that exhibit potent and selective antifungal activity (Fig. 2). Using chemoinformatic tools [2], these diverse chemicals can be grouped according to their chemical archetype, which can facilitate directed discovery efforts towards exotic scaffolds with uncommon modes of action (Fig. 2). Informatic navigation of antifungal chemical space could expedite the sampling of evolved scaffolds and enable researchers to hone in on active chemical moieties present in natural product extracts with superior precision, sensitivity, and efficiency to bioactivity-guided fractionation. By targeting antifungal natural products using our informatic

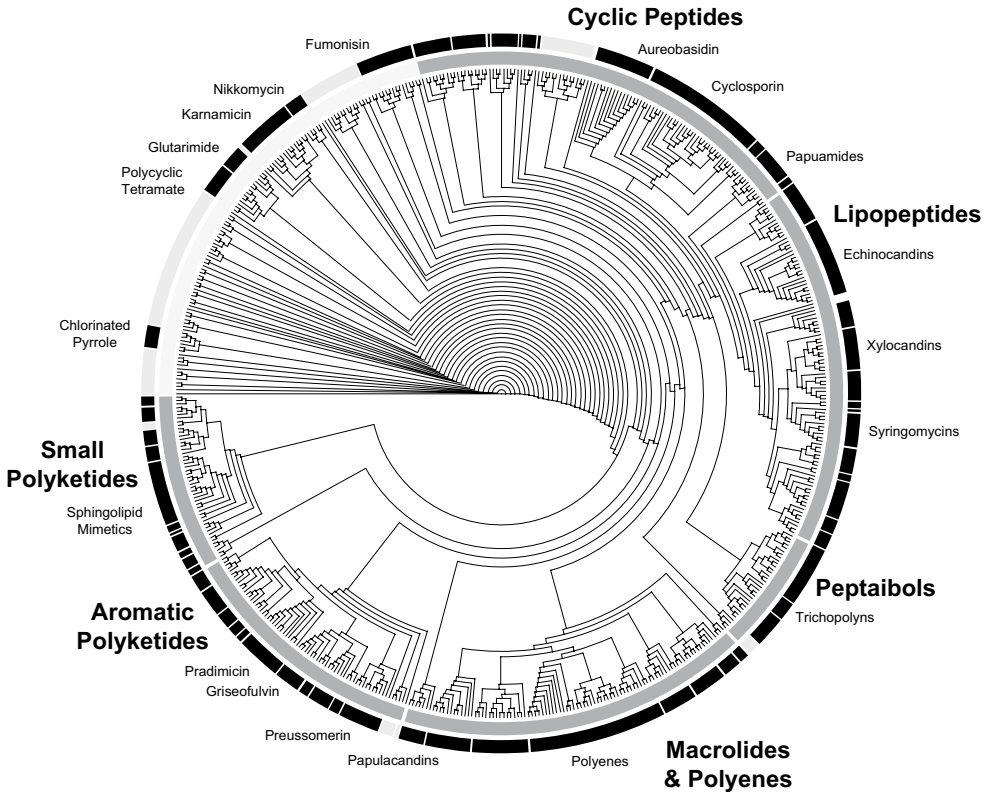

Fig. 2 Antifungal microbial natural products are a structurally diverse group of bioactive compounds. A chemoinformatic tree of known antifungal polyketides and nonribosomal peptides was constructed with ChemMine Tools, incorporating a comprehensive review of known natural products with antifungal activity. Structural similarity enables the identification of several super-groups, including lipopeptides, aromatic polyketides, and a selection of the peptaibols. Super-groups (*inner ring*; *dark grey*) are divided into sub-groups (*outer ring*), comprising nearly 100 unique chemical scaffolds (*outer ring*; *black*), such as cyclosporins, xylocandins, and trichopolyns

strategies, we hope to define an approach to discover and develop these crucial bioactive molecules.

Materials and methods

General experimental procedures

LCMS data were collected using a Bruker AmazonX ion trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, using a Luna C_{18} column (50, 150, or 250 × 4.6 mm, Phenomenex) for analytical separations, running acetonitrile and water with 0.1 % formic acid as the mobile phase. For analytical flow rates a UV/MS flow splitter of 10:1 was used. LCMS spectral analysis was performed using Compass DataAnalysis 4.1 (Bruker).

Strains and culture conditions

The environmental isolate of *Hypocrea minutispora* RKDO-344 and unidentified fungal isolate GE 174 were obtained from Great Slave Lake, Northwest Territories, Canada. *H. minutispora* RKDO-344 and GE 174 were maintained on SMYA agar (10 g/L peptone, 40 g/L maltose, 10 g/L yeast extract, 15 g/L agar) at 22 °C.

Hierarchical clustering of antifungal natural products

A comprehensive database of over 500 antifungal polyketides and nonribosomal peptide structures was assembled manually from the Dictionary of Natural Products and Antibase, as well from as a comprehensive natural products literature review. Structures were drawn manually or collected from ChemSpider or PubChem. Hierarchical clustering was performed using the hierarchical clustering function of ChemMine Tools, generating a newick tree that could then be displayed and analysed using Dendroscope, and subsequently labelled with Adobe Illustrator CS6.

Production, extraction, and detection of trichopolyn

For production of trichopolyn, RKDO-344 was inoculated from a 5-day shaking culture in SMYA media into MMK2 media (40 g/L mannitol, 5 g/L yeast extract, 4.3 g/L

Murashuge and Skoog salts) and grown standing at 22 °C at a 20 degree angle. Cultures were extracted with 5 % XAD7 and 5 % HP20 activated resins. For iSNAP analysis, trichopolyns were separated using a 150 mm Luna C_{18} column with a flow rate of 1 mL/min, ramping from 2 % acetonitrile at 5 min to 100 % acetonitrile at 25 min (curve 7). Trichopolyns 1, 2, and 3 were identified from environmental extract 344-M3 using the iSNAP algorithm with standard conditions, but with P1/P2 score cut-offs of 30/30 and with a 2 % noise filtering cut-off.

GNP trichopolyn variant identification

A structure database of α-aminoisobutyric acid (Aib) and alanine combinations, acyl tail length variations, and 2-amino-6-hydroxy-4-methyl-8-oxodecanoic acid oxidation variations was created for the trichopolyn scaffold based on trichopolyn 1. In addition, all structural combinations of valine, isoleucine, and α-aminoisobutyric acid were also included to afford a final structural database consisting of 432 compounds. This combinatorial database was created through the GNP online interface (http://magarveylab.ca/gnp/). Structural confirmation was carried out through manual MS2 annotation and iSNAP fragment hit analysis.

iSNAP analogue programme for trichopolyn variant identification

The LC–MS/MS data file of the RKDO-344 extract containing trichopolyns was analysed with the iSNAP analogue programme through the online interface (http://magarveylab.ca/analogue/) in precise search and analogue search mode. Precise search settings were identical to those described earlier, while analogue search settings used the standard minimum mass difference (12 Da) and a maximum mass difference of 30 Da, with the known trichopolyns as defined seed structures. Structural confirmation was carried out through manual MS2 annotation and iSNAP fragment hit analysis.

Results and discussion

To demonstrate the utility of our informatic discovery platforms for identifying novel variants of targeted natural products, we screened a library of extracts of environmental fungi using LC–MS/MS. One of these samples from our library (RKDO-344) was an extract from an isolate of *Hypocrea minutispora* and was found by iSNAP dereplication analysis to contain an antifungal peptide family known as the trichopolyns [8]. These peptides are a class of linear antifungal peptides referred to as peptaibols [5], which represent a small fraction of the diverse classes of antifungal polyketides and nonribosomal peptides (Fig. 2), but are themselves a well-established family (Fig. S1). Repetitive α-aminoisobutyric acid (Aib) monomers frequently installed within the peptaibol peptide backbones typically leads to an alpha helical structure and directs membrane channel formation [3]. Peptaibols are known to have a nonribosomal origin [5], and like other nonribosomal peptides, variation occurs through tailoring modifications and nonspecific amino acid incorporation during their assembly [6], creating within-family diversity which may be exploited to identify more selective agents.

We chose to expand on the known chemical space associated with the trichopolyns by investigating this extract further using our two automated informatic approaches, GNP and the iSNAP analogue programme. Using our GNP methodology, we used the trichopolyn structure as a scaffold for combinatorialization [16], incorporating changes from promiscuous amino acid incorporation, as well as modifications observed in related peptaibols such as leucinostatin and trichoderin. This combinatorialization of structural elements yielded a library of 432 hypothetical structures that could result from the trichopolyn assembly line, based on seven sites of diversification in our trichopolyn scaffold. Next, we used a GNP-specific iSNAP natural product detection algorithm to process LC–MS/MS data and assess whether any predicted known or unknown trichopolyns from our library could be identified in our sample. Reanalysis of LC–MS/MS data of the trichopolyn extract with this extended hypothetical variant library revealed three known [8] and three novel structures (Fig. 3). The trichopolyn structures detected by iSNAP at retention times 27.18 (1), 27.50 (2), 27.58 (3), 28.06 (4), 28.51 (5), and 28.85 min (6) had MS/MS fragmentation patterns that matched the iSNAP-generated fragmentation patterns of the hypothetical variant structures #191 (1; trichopolyn 4), #192 (2; trichopolyn 2), #188 (3; trichopolyn 1), #240 (4; 1174 Da), #236 (5; 1188 Da), and #284 (6; 1190 Da). Manual MS/MS annotation confirmed the identity of each of these iSNAP-detected variants (Figs. S2–S7), demonstrating that this approach is useful and accurate in expanding and exploring chemical space around known natural product structures.

We chose to investigate this extract further using the distinct iSNAP analogue profiling tool, both to check for the presence of trichopolyns that may have been overlooked by GNP, and to demonstrate the differences between these informatic discovery methodologies. To analyse the LC–MS/MS data with the iSNAP analogue function, we loaded the trichopolyn extract data file into the iSNAP interface for precise and analogue search. In contrast to GNP, the iSNAP analogue function does not require the generation of a combinatorialized scaffold library. Instead, dereplicated structures or defined seeds are used to search for possible

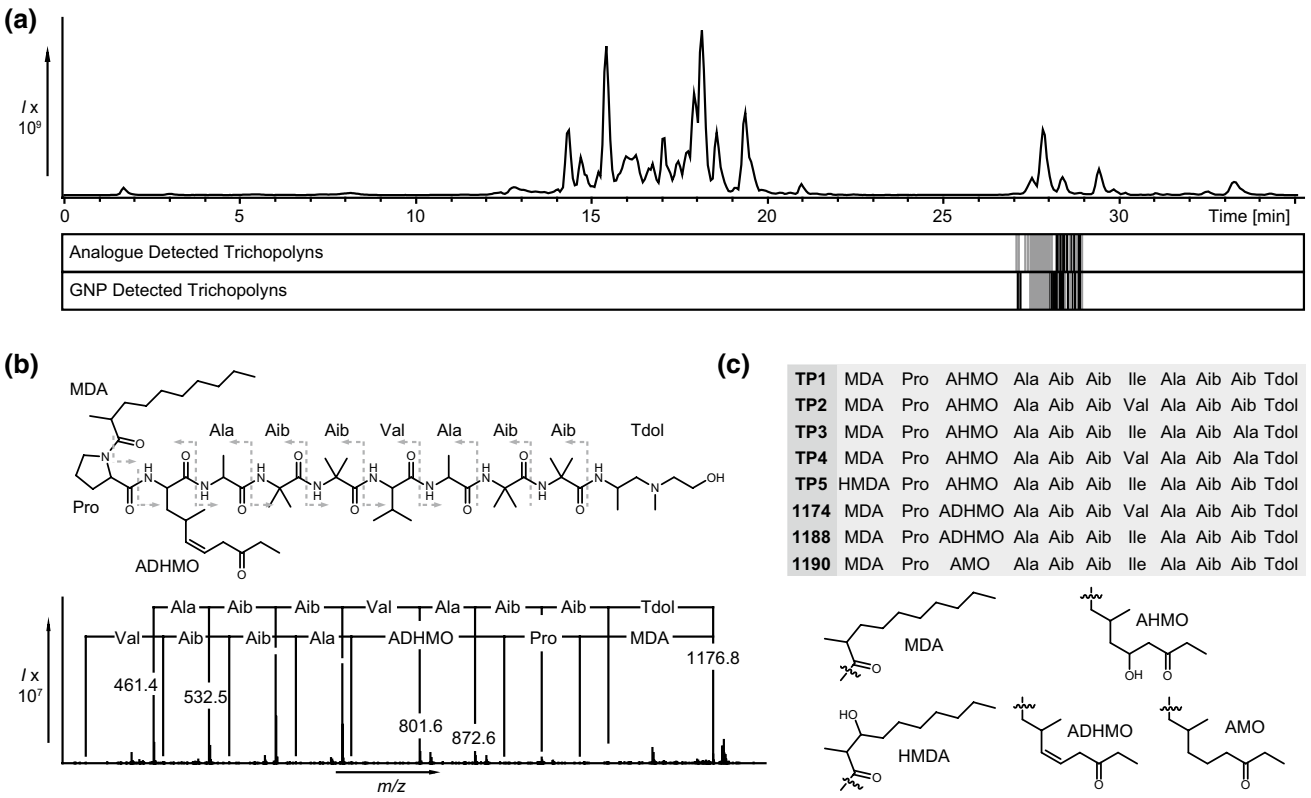

Fig. 3 Expansion of the trichopolyn family of natural products through automated analogue detection. **a** Processing a *Hypocrea minutispora* extract with LC–MS/MS and iSNAP led to the identification of the trichopolyn family of antifungal peptaibols, including trichopolyns 1-3 (*grey*; Figs. S2–S4). In an attempt to identify novel variants we utilised the iSNAP analogue function, and also generated a series of hypothetical variants based on variation observed in related peptaibols for analysis with GNP. Reanalysis with the iSNAP analogue function and GNP enabled detection of three novel trichopolyns (*black*; Figs. S5–S7). All scoring parameters are listed in *Methods*. **b** MS/MS sequencing of a proposed new trichopolyn variant (1174 Da). **c** Alignment of known and proposed new trichopolyns, including rare monomers such as 2-methyldecanoic acid (MDA), 2-methy-3-hydroxydecanoic acid (HMDA), 2-amino-6-hydroxy-4-methyl-8-oxodecanoic acid (AHMO), 2-amino-6-dehydro-4-methyl-8-oxodecanoic acid (ADHMO), 2-amino-4-methyl-8-oxodecanoic acid (AMO), α-aminoisobutyric acid (Aib), and trichodiaminol (Tdol)

analogues, localising differences between parent and analogue mass down to a single monomer using automated MS/MS analysis (Yang et al. resubmitted with revisions). Because of this, the iSNAP analogue programme is a more rapid means of analysing LC–MS/MS spectra and identifying potential analogues, but the diversity of analogues that can be accurately detected is more limited. Consistent with this observation, analysis of the trichopolyn extract with the iSNAP analogue programme revealed additional scans of trichopolyns 4, 2, and 1, identifying these both as dereplicated structures and as analogues through their respective site-specific modifications (Fig. 3). The analogue programme also detected most of the scans GNP had identified as trichopolyn variants, excluding the most minor variant (detected by GNP as hypothetical variant #240), which was only detected in a single MS/MS scan by GNP. Overall, these results demonstrate that both of these informatic techniques are useful tools to reveal new, minor variants of natural products from complex backgrounds.

Natural products are an important source of pharmaceutically relevant chemical scaffolds, particularly for antimicrobials [4, 18]. Although natural products have been the main source of antimicrobial agents since the discovery of penicillin, traditional bioactivity-guided approaches were laborious and demonstrated relatively low sensitivity, leading to the abandonment of industrial natural products discovery programmes [9, 11]. In light of recent advances in informatic technologies, and, in particular, the growing success of cheap and rapid microbial genome sequencing, natural products are again beginning to garner attention [1]. Rapid profiling of culture extracts using GNP and the iSNAP analogue programme can identify large families of valuable natural products with resolution that was impossible to achieve with bioactivity-guided isolation methods, demonstrated here with our mining of the trichopolyns. GNP and the iSNAP analogue programme offer unique opportunities to explore microbial chemistry, as both are automated platforms that are easily accessible through user-friendly

interfaces (http://magarveylab.ca/gnp/ and http://magarveylab.ca/analogue/) and do not require high-resolution MS data to facilitate the discovery of new compounds. Contemporary approaches such as molecular networking [17] can also provide a means of locating natural product variants, but require extensive manual MS/MS annotation to reveal structural alterations, and require direct comparisons with known compounds [19] or genetic knock-outs [17] to determine chemical identity and biosynthetic origin. More automated approaches such as CycloQuest [14], NRPQuest [13], and RippQuest [12] can facilitate a streamlined means of detecting natural products, but are limited in their scope and efficacy relative to GNP or the iSNAP analogue programme, representing specialised programmes for specific classes of natural products. At the moment, GNP and the iSNAP analogue programme present a unique opportunity for advancing natural products discovery programmes, allowing researchers to significantly expedite the detection of natural product congers toward new structures—and potentially—improved activities. By continuing targeted discovery efforts like those outlined in this work, we hope to expedite the elaboration and exploration of antifungal chemical scaffolds to address clinical need.

Acknowledgments This work was funded through a Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grant (RGPIN 371576-2014) (NAM) and a Joint Programme Initiative on Antimicrobial Resistance funded through the Canadian Institutes of Health Research (NAM). CWJ is funded through a CIHR Doctoral Research Award. NAM is supported by the Canada Research Chairs Program.

References

1. Bachmann BO, Van Lanen SG, Baltz RH (2014) Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? J Ind Microbiol Biotechnol 41(2):175–184. doi:10.1007/s10295-013-1389-9
2. Backman TW, Cao Y, Girke T (2011) ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res 39(Web Server issue):W486W–491. doi:10.1093/nar/gkr320
3. Chugh JK, Wallace BA (2001) Peptaibols: models for ion channels. Biochem Soc Trans 29(Pt 4):565–570
4. Clardy J, Fischbach MA, Walsh CT (2006) New antibiotics from bacterial natural products. Nat Biotechnol 24(12):1541–1550. doi:10.1038/nbt1266
5. Degenkolb T, Kirschbaum J, Brückner H (2007) New sequences, constituents, and producers of peptaibiotics: an updated review. Chem Biodivers 4(6):1052–1067. doi:10.1002/cbdv.200790096
6. Fischbach MA, Clardy J (2007) One pathway, many products. Nat Chem Biol 3(7):353–355. doi:10.1038/nchembio0707-353
7. Ibrahim A, Yang L, Johnston C, Liu X, Ma B, Magarvey NA (2012) Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc Natl Acad Sci USA 109(47):19196–19201. doi:10.1073/pnas.1206376109
8. Iida A, Mihara T, Fujita T, Takaishi Y (1999) Peptidic immunosuppressants from the fungus *Trichoderma polysporum*. Bioorg Med Chem Lett 9(24):3393–3396. doi:10.1016/S0960-894X(99)00621-6
9. Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. Nat Rev Drug Discov 4(3):206–220. doi:10.1038/nrd1657
10. Lam KS (2007) New aspects of natural products in drug discovery. Trends Microbiol 15(6):279–289. doi:10.1016/j.tim.2007.04.001
11. Li JW, Vederas JC (2009) Drug discovery and natural products: end of an era or an endless frontier? Science 325(5937):161–165. doi:10.1126/science.1168243
12. Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, Brewer HM, Pasa-Tolic L, Bandeira N, Moore BS, Pevzner PA, Dorrestein PC (2014) Automated genome mining of ribosomal peptide natural products. ACS Chem Biol 9(7):1545–1551. doi:10.1021/cb500199h
13. Mohimani H, Liu WT, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA (2014) NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. J Nat Prod 77(8):1902–1909. doi:10.1021/np500370c
14. Mohimani H, Liu WT, Mylne JS, Poth AG, Colgrave ML, Tran D, Selsted ME, Dorrestein PC, Pevzner PA (2011) Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. J Proteome Res 10(10):4505–4512. doi:10.1021/pr200323a
15. Ostrosky-Zeichner L, Casadevall A, Galgiani JN, Odds FC, Rex JH (2010) An insight into the antifungal pipeline: selected new molecules and beyond. Nat Rev Drug Discov 9(9):719–727. doi:10.1038/nrd3074
16. Schüller A, Hähnke V, Schneider G (2007) SmiLib v2.0: a Java-Based Tool for Rapid Combinatorial Library Enumeration. QSAR Comb Sci 26(3):407–410. doi:10.1002/qsar.200630101
17. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC (2012) Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci USA 109(26):E1743–E1752. doi:10.1073/pnas.1203689109
18. Wright GD (2012) Antibiotics: a new hope. Chem Biol 19(1):3–10. doi:10.1016/j.chembiol.2011.10.019
19. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, Glukhov E, Wodtke A, de Felicio R, Fenner A, Wong WR, Linington RG, Zhang L, Debonsi HM, Gerwick WH, Dorrestein PC (2013) Molecular networking as a dereplication strategy. J Nat Prod 76(9):1686–1699. doi:10.1021/np400413s