7.13 Uncovering Bioactive Natural Products Via Biochemometric Methodologies

Joshua J Kellogg and Nadja B Cech, Department of Chemistry & Biochemistry, University of North Carolina at Greensboro, Greensboro, NC, United States

© 2020 Published by Elsevier Ltd. All rights reserved.

7.13.1 Introduction

Natural products represent a tremendously rich resource for drug discovery paradigms or as dietary supplements to improve human health. In their unrefined form, natural products consist of complex mixtures of many structurally diverse components. The identities of bioactive constituents are (a) often not known, and (b) likely to differ depending on the biological activity evaluated. The central challenge facing natural product researchers is unraveling the complexities of such complex mixtures and assigning structures and biological activities to the components. With this chapter, we review various strategies that can be employed towards this goal and discuss their advantages and disadvantages. We summarize the traditional approach of bioassay guided fractionation and describe how newer methodologies (biochemometrics strategies) that integrate metabolite profiles (metabolomics data) with biological activity evaluation can be employed to enhance the bioassay guided fractionation process.

7.13.2 Bioassay-Guided Fractionation

The traditional approach to discovering bioactive constituents from complex natural product mixtures is a range of related techniques generally referred to as "bioassay-guided fractionation." With bioassay guided fractionation, extracts are simplified and separated using various chromatographic techniques, the resulting (sub)fractions are screened for biological activity, and the process is repeated iteratively until a single bioactive constituent has been isolated and identified (Fig. 1). Bioassay-guided fractionation has remained the "gold standard" of natural product discovery for the last few decades, and has resulted in the discovery of many bioactive compounds, including paclitaxel, artemisinin, and vincristine/vinblastine.

Despite its significance in natural product discovery, it is not uncommon with bioassay guided fraction to experience a loss of activity or failure to isolate a single bioactive constituent from the complex mixture. The bioactive principles are sometimes minor constituents, and bioassay-guided fractionation tends to be biased towards the most abundant (or highly responsive) components in a mixture. This bias has also led to instances of re-isolation of already known active compounds. The problem of re-isolation of known compounds is typically addressed using various "dereplication" approaches, which often rely on comparing spectroscopic data against libraries and databases containing the experimental data of known compounds.

7.13.3 A Note on In Vitro Data

Any attempt to identify or characterize bioactive principles from a complex mixture axiomatically requires biological evaluation of the extract(s), fraction(s), or isolated compounds. The most common assessment of biological activity is an in vitro assay that analyzes cellular response to a stimulus, or inhibition/activation of an isolated protein/enzyme. In vitro assays have been used to evaluate complex mixtures for potential lead compounds to address an array of health conditions, including chronic diseases such as diabetes, cancer, and obesity, infectious diseases (e.g., bacterial, fungal, and parasitic infections), and to promote overall health and wellness. Cell-based assays have been used to provide insight into potential interactions between natural products and conventional pharmaceuticals, investigate how natural products interact with the gut microbiome, and to model how nutritional matrices could impact the pharmacodynamics and pharmacokinetics of natural product constituents. Furthermore, in vitro data have the potential to provide mechanistic or structure-function insight, which can be imperative in guiding future evaluation via mathematical or statistical modeling experiments, as well as in vivo or clinical studies.

Current address: Department of Veterinary and Biomedical Sciences, Pennsylvania State University, University Park, PA, United States.
However, it is worth noting that in vitro assay data does not always translate to activity in the whole organism. This is due to the increased complexity of whole biological systems (i.e., organisms vs single cells/isolated proteins), and the potential effects of other biological factors (e.g., cytosol composition, microbiota, etc.) that are not accounted for in most in vitro studies. García-Contreras et al. found that reaction medium composition plays a role in enzyme activity, and thus researchers must take care extrapolating conclusions from in vitro studies investigating enzymatic inhibition by natural products, as there could be augmentation of activity and efficacy when scaling up to a more complex model system. Another challenge is that many of the purported health benefits of botanical natural products (e.g., "maintain joint health and flexibility", "support proper immune function", or "enhance adaptability to stress") are difficult to model with an in vitro assay. Successful bioassay-guided fractionation of natural products mixtures necessitates the availability of a relevant, robust and translatable biological assay.

7.13.4 Biochemometrics Modeling

Metabolomics represents an inclusive approach to analyze the entirety of measure metabolites from a complex mixture and is a rapidly emerging field in a number of disciplines, including natural product chemistry and drug discovery. Metabolomics studies can utilize data from a variety of analytical instrumentation, though mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) are the most common. Metabolomic profiling results in the generation of large datasets comprising potentially major and minor components from the mixture and require tools to parse the data and provide meaningful analysis. The most commonly employed statistical tool for comparing chemical composition of complex mixtures is principal component analysis (PCA). However, PCA analysis only takes into consideration the chemical data. Biochemometrics has been successfully applied by different groups to identify bioactive principles from natural product mixtures. For example, a recent study investigating the anti-tuberculosis (Mycobacterium tuberculosis H37Rv) activity of volatile
components the essential oil of *Eucalyptus citriodora* used Pearson correlation to identify 37 volatile anti-tuberculosis compounds (Fig. 2). The monoterpene citronellol demonstrated significant inhibition of tuberculosis in the airborne stage, and the observed activity was attributed to mixtures of several compounds (citronellol, (S)-citronella and eucalyptol).57

Partial least squares (PLS) has become one of the dominant statistical modeling approaches for biochemometrics studies. Biochemometric analyses are frequently multi-dimensional models, and can be difficult to decipher. To this end, a number of visualization techniques have been developed to interpret PLS models, with the S-plot and the selectivity ratio being the two dominant metrics.43

Abdelmohsen et al. (2014) employed OPLS-DA to model potential bioactive metabolites from *Actinokineospora* sp. strain EG49, an actinomycete cultivated from the sponge *Spheciospongia vagabunda*. Using the source of the metabolite (broth supernatant vs. cell) as the dependent variables, the model generated an S-plot (Fig. 3), in which the upper and lower extremes reveal metabolites that most strongly correlate and covary with the biological activity, suggesting they are responsible for the observed activity. This study yielded two new O-glycosylated angucyclines, named actinosporins A and B, which were isolated and subsequently demonstrated activity against *Trypanosoma brucei brucei*.58 S-plots were also used to identify a neuroprotective agent against Parkinson’s disease. The S-plot, from an OPLS model of *Alpinia oxyphylla* fractions and associated neuroprotective activities, revealed oxyphylla A as a potentially powerful neuroprotectant.59 Finally, studying the stipe of *Ganoderma sinense*, the S-plot approach was used to uncover five cytotoxic compounds to prioritize for follow-up studies.53

A complimentary approach for highlighting significant PLS model features from the independent dataset is the selectivity ratio, which provides a quantitative metric for each variable and its ability to distinguish between different groups. The selectivity ratio is defined as the ratio between the explained and residual variance of the spectral variables (features) of a target-projection from the PLS model. With selectivity ratio modeling, features (spectral variables) that have high selectivity ratio have are more strongly associated with bioactive (rather than nonbioactive) samples.43 A recent study compared the use of the S-plot and selectivity ratio analysis when PLS datasets were employed to identify antimicrobial constituents from two endophytic fungi (*Alternaria* sp. and *Pyrenochaeta* sp.). For *Pyrenochaeta* sp., both the S-plot and selectivity ratio identified macrospelhide A as the dominant antimicrobial component. However, interpreting the PLS model from *Alternaria* sp., the S-plot suggested alternariol monomethyl ether as the main bioactive principle, while the selectivity ratio yielded altersetin as the responsible compound. Follow-up assaying confirmed altersetin as the more potent bioactive component, with a minimum inhibitory concentration of 0.59 μM against a sensitive strain of *Staphylococcus aureus* (compared to 275 μM for alternariol monomethyl ether).61 In a pharmacokinetic study of *Camellia sinensis* (green tea), a PLS model was used to predict which constituents were responsible for the in vitro inhibition of intestinal UDP-glucuronosyltransferases (UGTs). The selectivity ratio predicted (−)-epicatechin gallate (ECG) as a major contributor to inhibition of UGTs (Fig. 4).62 This prediction was validated with follow up assays of isolated ECG.
Fig. 3  Orthogonal partial least square-discriminant analysis (OPLS-DA) (A) and resulting S-plot (B) of supernatant broth (red) vs. cellular (blue) extracts from cultures of Actinokineospora sp. strain EG49. The S-plot’s upper (blue dots) and lower (red dots) termini represent unique secondary metabolites that differentiated the two classes, and was used to identify actinosporins A and B. Reproduced with permission from Abdelmohsen, U. R.; Cheng, C.; Viegelmann, C.; Zhang, T.; Grkovic, T.; Ahmed, S.; Quinn, R. J.; Hentschel, U.; Edrada-Ebel, R. Dereplication Strategies for Targeted Isolation of New Antitrypanosomal Actinosporins A and B From a Marine Sponge Associated-Actinokineospora sp. EG49. Mar. Drugs 2014, 12 (3), 1220, MDPI.

Fig. 4  Biochemometric analysis of the inhibition of UDP-glucuronosyltransferase (UGT) activity by green tea constituents. (A) Mass spectrometry-based metabolomics was correlated with bioactivity to generate a partial least squares (PLS) model, and the selectivity ratio for each feature detected in the green tea extract. The more negative values on the y-axis represent more significant contribution to the observed UGT inhibitory activity. (B) All six detected ions represented isotope peaks and adduct ions of (−)-epicatechin gallate (ECG) molecular ion. Reproduced with permission from Tian, D.-D.; Kellogg, J. J.; Okut, N.; Oberlies, N. H.; Cech, N. B.; Shen, D. D.; McCune, J. S.; Paine, M. F. Identification of Intestinal UDP-Glucuronosyltransferase Inhibitors in Green Tea (Camellia sinensis) Using a Biochemometric Approach: Application to Raloxifene as a Test Drug Via In Vitro to In Vivo Extrapolation. Drug Metab. Dispos. 2018, 46, 552–560, American Society for Pharmacology and Experimental Therapeutics (ASPET).
7.13.5 Structural Annotation

Biochemometric methodologies rely on the “raw” spectral data provided by the analytical instrumentation as independent variable input. For mass spectrometry-based metabolomics, these data take the form of the \( m/z \) value of the analyte, which can be paired with retention time if an LC-MS system is employed. For NMR metabolomics, the spectroscopic variables represent binned peak signals for the nucleus of interest (traditionally \(^1\text{H}\) or \(^{13}\text{C}\)) and depends on the resolution of the acquisition instrument.\(^{63-65}\) However, this does not provide any structural information about the putative bioactive metabolites, especially if they are unknown.

Attempts to better annotate and elucidate chemical structural information combined biochemometrics analyses with metabolomics tools that provide structural insight into the potential bioactives. One such approach utilizes the Global Natural Products Social Molecular Networking (GNPS) approach, in which MS/MS fragmentation data is mined from metabolomics datasets to provide a network of structurally-similar metabolites.\(^{16}\) GNPS has proven to be a powerful tool in annotation and de-replication of mass spectrometry-based datasets.\(^{66,67}\) For example, a recent study combined biochemometrics analysis of Angelica keiskei with molecular networking to tentatively annotate putative bioactive metabolites (Fig. 5). The analysis identified a set of chalcone analogs as being potentially active, and subsequent isolation efforts yielded two known components as well as one compound that

![Figure 5](image-url)
Fig. 6  A representative natural product extraction and fractionation scheme. The sample material first undergoes solvent extraction, after which it is subjected to liquid-liquid partitioning to produce an organic extract. This residue is solubilized and then fractionated with a normal-phase or open column chromatography system, and if further purification is necessary, the fractions are separated via reverse-phase HPLC to yield purified compounds. Reproduced with permission from Kellogg, J. J.; Paine, M. F.; McCune, J. S.; Oberlies, N. H.; Cech, N. B. Selection and Characterization of Botanical Natural Products for Research Studies: A NaPDI Center Recommended Approach. Nat. Prod. Rep. 2019, 36 (8), 1196–1221, Royal Society of Chemistry.
had no previously reported antimicrobial activity. The GNPS approach has been adapted further to create “bioactive molecular networking,” where the two analyses are harmonized to directly visualize bioactive metabolites in the context of their structural relationship to one another. Nothias et al. developed a biochemometric workflow to map a “bioactivity score” within a molecular network, and applied this approach to discover antiviral compounds from *Euphorbia dendroides*. This was achieved after a traditional bioassay-guided fractionation procedure was unable to discover the bioactive candidate molecules.

### 7.13.6 Confirmation of Activity

Biochemometric modeling and interpretation is a statistical approach to prioritize features as putative bioactive metabolites from a more complex mixture. While efforts to identify and quantify constituents of natural product extracts often rely on purified standards, this approach is viable only when standards are available. The alternative, when standards of the compounds of interest are not commercially available, is proceeding with an isolation scheme for the compounds from the complex starting material. A representative review of isolation procedures and techniques is available elsewhere. Generally, a common approach for natural product separation and isolation of constituents is a combination of liquid-liquid partitioning and column chromatography approaches (Fig. 6). A natural product is macerated with an organic solvent, typically methanol, and then partitioned between an immiscible binary solvent system (e.g., chloroform and water or ethyl acetate/water), followed by a “de-fatting” step using hexane-organic solvent partitioning to remove lipid-soluble components. This methodology is inherently flexible and could easily be modified with acid/base conditions or substituting other solvents as the system demands. The extracted residue is dried (typically under nitrogen or reduced pressure) and undergoes normal phase or open column chromatography to provide a first round of fractions. If necessary, fractions are subfractionated using one or more rounds of reversed phase preparative HPLC. This can be a laborious process, and the isolation of even milligrams quantities of a pure isolated compound could take weeks to months of work, and it is possible that it may be ultimately unsuccessful. This resource- and time-intensive is another argument in favor of using biochemometric approaches to guide isolation efforts.

### 7.13.7 Discussion

The identification of bioactive constituents from complex mixtures, with minimal reliance on multiple bioactivity-guided isolation steps remains a challenge for natural product discovery programs. The adaptation of statistical models to integrate metabolomic datasets with corresponding in vitro assay data, represents an important step to improve the efficiency and productivity of discovery efforts. And the combination of complimentary methods (e.g. biochemometrics and molecular networking) has the potential to provide new tools to accelerate discovery. As both the analytical and statistical capabilities of researchers improve, these techniques will drive secondary metabolite discovery and application, and new arenas for combining orthogonal datasets will aid in identifying biological roles of these small molecules.

### References

Uncovering Bioactive Natural Products Via Biochemometric Methodologies


