COMMUNICATION

OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment†

Kathleen R. Murphy,*[ab] Colin A. Stedmon,[c] Philip Wenig[d] and Rasmus Bro[e]

An online repository of published organic fluorescence spectra has been developed, which can be searched for quantitative matches with any set of unknown spectra. It fills a critical gap by increasing access to measured and modelled (PARAFAC) spectra, and linking across studies and systems to reveal "global" fluorescence trends.

Fluorescence spectroscopy offers an inexpensive, non-destructive method for obtaining sensitive measurements of a diverse group of organic compounds that contain fluorophores. This technology is now widely used to characterise naturally-occurring organic matter in natural and artificial aquatic systems with the purpose of understanding how the fluorescent fraction of carbon is partitioned between different organic matter fractions, and inferring the processes responsible for its formation and removal.[1–5] With Excitation–Emission Matrix (EEM) spectroscopy, fluorescence emission is measured over a range of excitation wavelengths to produce three-dimensional fluorescence landscapes (Fig. 1). Each EEM represents total fluorescence from an unknown number of underlying fluorophores which in ideal conditions fluoresce independently following Beers Law, but under non-ideal conditions may interact.[6] Over the past ten years it has become common practice to decompose EEM datasets mathematically using PARAllel FACtor analysis (PARAFAC).[7–9] PARAFAC reduces the EEM dataset into a small number of building blocks – referred to as 'underlying components' – each with a characteristic excitation and emission spectrum (Fig. 1). Each EEM in a dataset is modelled by a simple recipe in which the same building blocks are combined in varying amounts, reflecting their variable concentrations.

There are now well over 100 published PARAFAC models of dissolved and natural organic matter (both referred to hereafter as NOM) and over 500 published PARAFAC components.[9,10]

However, no agreed measure exists for determining whether the same PARAFAC components were found in different studies. Furthermore, while scientists have some idea of the chemical structures likely to be responsible for NOM fluorescence, few reference data are readily available and even fewer studies have drawn reliable comparisons between PARAFAC components and pure organic compounds. It is presently unclear how often PARAFAC components extracted from NOM accurately represent the spectra of pure compounds or mixtures, or the degree to which PARAFAC decompositions are impaired by potential non-ideal chemical behaviours such as spectral shifting,[4] energy or electron transfer,[6,11] and charge–transfer interactions.[12]
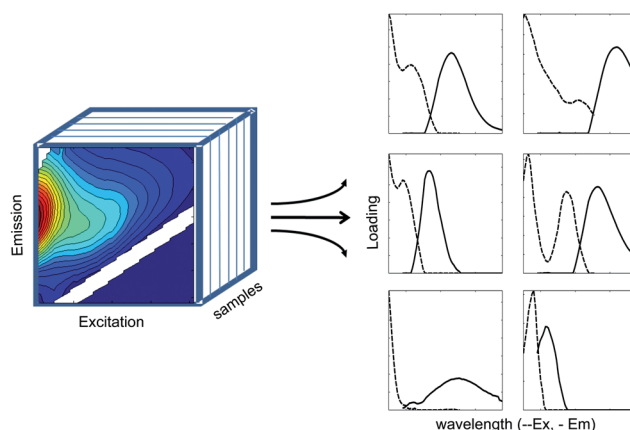
[a]University of New South Wales, Water Research Centre, Sydney, Australia. E-mail: krm@unsw.edu.au; Fax: +61 2 9313 8624; Tel: +61 2 9385 4601

[b]Chalmers University of Technology, Water Environment Technology, Gothenburg, Sweden

[c]Technical University of Denmark, National Institute for Aquatic Resources, Charlottenlund, Denmark. E-mail: cost@aqua.dtu.dk

[d]Ernst-Kabel-Stieg 5a, 22087 Hamburg, Germany. E-mail: philip.wenig@gmx.net

[e]University of Copenhagen, Dept. Food Science, Frederiksberg, Denmark. E-mail: rb@life.ku.dk

† Electronic supplementary information (ESI) available: Source of PARAFAC components depicted in Fig. 2. See DOI: 10.1039/c3ay41935e

Fig. 1 A dataset of fluorescence excitation emission matrices (EEMs) decomposed into six underlying components using PARAFAC.

It is widely supposed that spectrally similar PARAFAC components extracted from unrelated datasets are attributable to similar organic matter sources, and depict the same or similar underlying compounds having similar ecological functions. However, since the spectra of published PARAFAC components are only typically available as images or summary tables in the original publications, this hypothesis is extremely difficult to test. Thus, Ishii and Boyer[13] recently reviewed the reported distributions and responses to physicochemical processes of three apparently widespread humic-like PARAFAC components, finding numerous inconsistencies between studies with regard to their reported behaviours. However, in that review as in the overwhelming majority of reviewed studies, PARAFAC components were equated on the basis of broad criteria such as the number and positions of spectral peaks, with peak positions approximately defined and allowed to vary over a broad wavelength range. Previously, In the literature, PARAFAC components have been equated to specific compounds and redox states with little or no quantification of spectral similarity. This widespread use of qualitative or subjective criteria for equating components between studies is a serious confounding factor for interpreting global trends in component distributions and behaviours, or for deducing the organic structures likely to be responsible for the observed patterns. Recent papers have emphasised the importance of standardised approaches to measuring EEMs[14,15] and deriving PARAFAC models,[9,16] and a systematic way of comparing the results of different studies is urgently needed.

To support quantitative comparisons of fluorescence spectra between studies, an open-access spectral database (http://www.openfluor.org) has been developed. The database is accessible using any modern web browser (e.g. Mozilla, Chrome, Internet Explorer) on desktops, tablets or smartphones. All interactions between the user and the database occur via a simple graphical user interface with no programming necessary. The supporting use of HTML5, jQuery and JavaScript create a rich and interactive graphical user interface within the browser. When search query is implemented on an unknown set of reference spectra, quantitatively similar spectra are retrieved from the database.

Algorithms for quantifying spectral similarity have been the subject of extensive research in other branches of analytical chemistry,[17–19] but are undeveloped in the context of fluorescence. Currently, OpenFluor identifies similar spectra as having Tucker congruence[20] θ exceeding 0.95 on the excitation and emission spectra simultaneously (eqn (1)). A more targeted search for matching spectra will be implemented in the future as improved algorithms for matching spectra become available.

$$\theta = \theta_{ex} \times \theta_{em} \geq 0.95 \qquad (1)$$

Records in the OpenFluor database are accompanied by synopses of the study that generated the data, including a short methodological description and an active link to the published record at http://dx.doi.org. Unregistered visitors to the website may temporarily upload spectra and search for quantitatively similar spectra in the database. Completion of a free one-time

registration process allows the user to browse descriptions of matching models, generate plots, and download matched data. Registered users may elect to submit published spectra to the database, thereby making their own research results available for searching by other members of the fluorescence community.

Fig. 2 illustrates the potential for a spectral database to reveal similarities as well as differences between PARAFAC spectra. Each of the humic-like components depicted in Fig. 2A–C fulfil the description of "reoccurring Component 2" described by Ishii and Boyer[13] (excitation maxima approximately <240–275 nm and 339–420 nm; emission approximately 434–520 nm). Dozens of other spectra in the OpenFluor database also conform to this general description, yet are relatively poor quantitative matches for these spectra. In Fig. 2B, the four PARAFAC components shown share nearly identical emission spectra, but the excitation spectra fall into two distinct groups, corresponding with datasets from water treatment plants in Denmark[21] and Australia[5] which have different excitation spectra than in the models of datasets from the Florida Everglades[22] and the South Atlantic Bight.[23] Since nearly all components have primary excitation maxima near the limits of the measured or modelled range (<250 nm), they are mainly distinguishable by the position of their secondary excitation peak in conjunction with the position of the emission maximum ($C_{ex/em}$). In Fig. 2C, the strongly overlapping components shown appear to mainly cluster in two sets, described here as $C_{400/518}$ nm and $C_{380/500}$ nm. The ESI† lists published sources for components in Fig. 2.

Fig. 2D–F depict humic-like components identified in a number of studies, each fulfilling the description of Ishii and Boyer's reoccurring component 3 (excitation maxima approximately 240–260 nm and 295–380 nm, and emission maximum
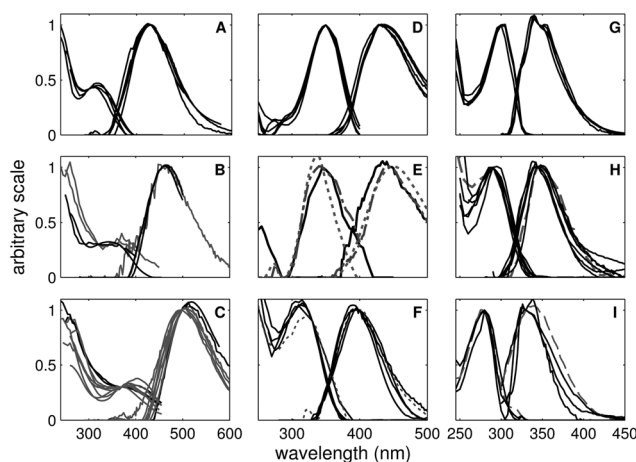

Fig. 2 Excitation (left) and emission (right) spectra of widely distributed PARAFAC components in the OpenFluor database. (A–F) are humic-like components and (G–I) are protein-like. (A) $C_{320/420}$; (B) $C_{370/460}$ and $C_{345/460}$; (C) $C_{400/518}$ nm and $C_{380/500}$ nm; (D) $C_{350/430}$ in five water treatment plant models from a single study; (E) qualitatively similar components to $C_{350/430}$ in other studies; (F) $C_{320/400}$; (G) $C_{300/340}$ displaying fine structure in the emission spectrum; (H) $C_{295/356}$ correlated to lignin in one study; and (I) $C_{275/350}$, similar to free dissolved tryptophan.

approximately 374–450 nm).[13] The component depicted in Fig 2D was identified repeatedly in a study of water treatment plants around Australia,[5] in which samples were measured on a single instrument but independent PARAFAC models were developed for each plant. A similar component is seen in several other studies (Fig 2E), although those spectra are more variable. Fig 2F depicts a different component, or given the apparent continuum of peak locations, possibly a suite of components representing different compounds or groups thereof. As the number of datasets in OpenFluor increases, a more robust picture of such components should emerge.

Fig. 2G–I illustrate three different protein-like components in the database that have each been described as "tryptophan-like". Fig. 2G depicts a component common to studies that sampled in Baltic[24] sea ice, Antarctic[25] sea ice, the North Atlantic ocean[24,26] and the Florida Everglades.[22] The spectra are extremely similar in each study, down to fine detail in the emission spectra, which suggests that a discrete organic compound rather than a mixture of compounds may be responsible for this signal. Fig. 2H depicts a component identified in models from natural and artificial environments.[5,27,28] The component depicted using dashed lines in this figure was strongly correlated with lignin concentration in one study.[27] Fig. 2I depicts a commonly-observed component with spectra similar to free dissolved tryptophan. The shape of the emission spectrum for this component differs between studies, possibly because it is derived from a group of compounds, and possibly also because interference by Raman scatter makes it difficult to accurately resolve its spectra.

The OpenFluor spectral database aims to address a serious deficiency affecting the current interpretation of NOM-PARAFAC models. Thus, although it is widely assumed that spectrally similar PARAFAC components identified in unrelated studies have similar sources and ecological functions, quantitative spectral comparisons have been implemented only rarely[5,10] and with respect to a small number of studies. At the same time, many studies have drawn conclusions about the origins and behaviours of various components on the basis of qualitative comparisons with earlier studies. It is therefore likely that inconsistencies between reported behaviours of similar PARAFAC components are at least partly attributable to the unintentional grouping of NOM components that are spectrally similar, yet chemically and behaviourally distinct.

It is also important to realise that many fluorophores could have very similar spectra, so identifying similar PARAFAC components in two different studies does not guarantee that the same compounds are responsible in both cases. Fig. 3 compares a PARAFAC component identified in the Mackenzie River plume[29] in northern Canada with the spectrum of pure dissolved sodium salicylate (C.A. Stedmon, unpublished data), a common pharmaceutical derived from wintergreen plants. Since the Mackenzie River watershed is mostly covered by virgin forests and wetlands and is minimally influenced by human activities,[30] a pharmaceutical source for this component can be ruled out. Instead, it is more likely to represent forest-derived phenolic compounds with very similar spectral characteristics to sodium salicylate. The database may therefore be more useful
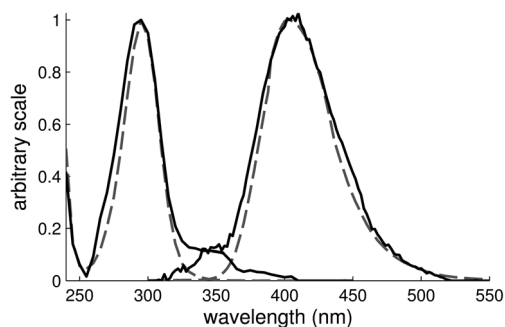

Fig. 3 Fluorescence spectra of a terrestrially derived PARAFAC component in the nearly pristine Mackenzie River watershed[29] (lines) compared with pure dissolved sodium salicylate (dashes).

for detecting patterns in the occurrence of fluorescence components, and deducing relationships between them, than as a tool for identifying the specific chemical structures responsible for the observed signals.

Conclusions

OpenFluor enables quantitative comparisons of fluorescence spectra between studies for the first time *via* a simple browser-based user interface. At release, the database contains over 200 PARAFAC spectra derived from more than 30 published studies of NOM in natural and industrial aquatic systems. Its size is expected to increase rapidly, since users can submit published spectra to the database *via* the online system in a matter of minutes, and doing so could greatly increase the chances that a study is encountered and cited by other researchers. Future developments to the database are planned to further increase its usefulness, including the incorporation of automated routines for checking the quality of fluorescence spectra, and the implementation of enhanced spectral-matching algorithms incorporating chemical as well as statistical criteria.

Acknowledgements

Notes and references

1 D. M. McKnight, E. W. Boyer, P. K. Westerhoff, P. T. Doran, T. Kulbe and D. T. Andersen, *Limnol. Oceanogr.*, 2001, **46**, 38–48.

2 K. R. Murphy, C. A. Stedmon, T. D. Waite and G. M. Ruiz, *Mar. Chem.*, 2008, **108**, 40–58.

3 C. A. Stedmon and S. Markager, *Limnol. Oceanogr.*, 2005, **50**, 1415–1426.

4 J. Ma, R. Del Vecchio, K. S. Golanoski, E. S. Boyle and N. V. Blough, *Environ. Sci. Technol.*, 2010, **44**, 5395–5402.

5 K. R. Murphy, A. Hambly, S. Singh, R. K. Henderson, A. Baker, R. Stuetz and S. J. Khan, *Environ. Sci. Technol.*, 2011, **45**, 2909–2916.

6 R. Del Vecchio and N. V. Blough, *Environ. Sci. Technol.*, 2004, **38**, 3885–3891.

7 R. Bro, *Chemom. Intell. Lab. Syst.*, 1997, **38**, 149–171.

8 C. A. Stedmon, S. Markager and R. Bro, *Mar. Chem.*, 2003, **82**, 239–254.

9 K. R. Murphy, C. A. Stedmon, D. Graeber and R. Bro, *Anal. Methods*, 2013, **5**, 6557–6566.

10 K. R. Murphy, R. Bro and C. A. Stedmon, in *Aquatic organic matter fluorescence*, ed. P. Coble, A. Baker, J. Lead, D. Reynolds and R. Spencer, Cambridge University Press, New York, (ISBN: 9780521152594), in press.

11 E. S. Boyle, N. Guerriero, A. Thiallet, R. Del Vecchio and N. V. Blough, *Environ. Sci. Technol.*, 2009, **43**, 2262–2268.

12 G. S. Furman and W. F. W. Lonsky, *J. Wood Chem. Technol.*, 1988, **8**, 165–189.

13 S. K. L. Ishii and T. H. Boyer, *Environ. Sci. Technol.*, 2012, **46**, 2006–2017.

14 D. Kothawala, K. Murphy, C. Stedmon, G. Weyhenmeyer and L. Tranvik, *Limnol. Oceanogr.: Methods*, 2013, **11**, 616–630.

15 K. R. Murphy, K. D. Butler, R. G. M. Spencer, C. A. Stedmon, J. R. Boehme and G. R. Aiken, *Environ. Sci. Technol.*, 2010, **44**, 9405–9412.

16 C. A. Stedmon and R. Bro, *Limnol. Oceanogr.: Methods*, 2008, **6**, 572–579.

17 H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold, *Proteomics*, 2007, **7**, 655–667.

18 S. E. Stein, *J. Am. Soc. Mass Spectrom.*, 1999, **10**, 770–781.

19 K. X. Wan, I. Vidavsky and M. L. Gross, *J. Am. Soc. Mass Spectrom.*, 2002, **13**, 85–88.

20 L. R. Tucker, A method for synthesis of factor analysis studies (Personnel Research Section Report no. 984), Department of the Army, Washington D.C., 1951.

21 C. A. Stedmon, B. Seredyńska-Sobecka, R. Boe-Hansen, N. Le Tallec, C. K. Waul and E. Arvin, *Wat. Res.*, 2011, **45**, 6030–6038.

22 M. L. Chen, R. M. Price, Y. Yamashita and R. Jaffe, *Appl. Geochem.*, 2010, **25**, 872–880.

23 P. Kowalczuk, M. J. Durako, H. Young, A. E. Kahn, W. J. Cooper and M. Gonsior, *Mar. Chem.*, 2009, **113**, 182–196.

24 C. A. Stedmon, D. N. Thomas, M. Granskog, H. Kaartokallio, S. Papadimitriou and H. Kuosa, *Environ. Sci. Technol.*, 2007, **41**, 7273–7279.

25 C. A. Stedmon, D. N. Thomas, S. Papadimitriou, M. A. Granskog and G. Dieckmann, *J. Geophys. Res.*, 2011, **116**, 1–9.

26 K. R. Murphy, G. M. Ruiz, W. T. M. Dunsmuir and T. D. Waite, *Environ. Sci. Technol.*, 2006, **40**, 2357–2362.

27 C. L. Osburn and C. A. Stedmon, *Mar. Chem.*, 2011, **126**, 281–294.

28 Y. Yamashita, J. N. Boyer and R. Jaffé, *Cont. Shelf Res.*, 2013, **66**, 136–144.

29 S. A. Walker, R. M. W. Amon, C. Stedmon, S. Duan and P. Louchouarn, *J. Geophys. Res.: Biogeosci.*, 2009, **114**, G00F06.

30 K. A. Morgenstern, D. Donahue and N. Toth, McKenzie River watershed baseline monitoring report 2000 to 2009, 2013, http://www.eweb.org/public/documents/water/baselineReportJan2011.pdf, accessed 29 October 2013.