Parallel factor analysis of spider fluorophores

Scott M. Reed\textsuperscript{a,*}, My-Trinh Do\textsuperscript{a}, Susan E. Masta\textsuperscript{b}

\textsuperscript{a}Department of Chemistry, Portland State University, P.O. Box 751, Portland, OR 97207, USA
\textsuperscript{b}Department of Biology, Portland State University, P.O. Box 751, Portland, OR 97207, USA

1. Introduction

Recent reports of fluorescence in parrots,\textsuperscript{[1]} mantis shrimp,\textsuperscript{[2]} flowers,\textsuperscript{[3]} termites,\textsuperscript{[4]} bees,\textsuperscript{[5]} and spiders\textsuperscript{[6,7]} greatly expand the list of organisms known to fluoresce. While fluorescent proteins in corals and jellies have been well characterized,\textsuperscript{[8,9]} the list of organisms known to fluoresce.

While fluorescent molecules have been isolated and identified from a number of organisms, the process of isolation is challenging and many biological fluorophores remain to be identified. There are several obstacles to isolating pure fluorophores and the process of isolation can alter the optical properties of the molecules from those observed in vivo. The local environment surrounding a fluorophore in vivo can change a molecule’s degree of protonation and can alter the availability of hydrogen-bonding or energy transfer partners that in turn can change the observable fluorescence. For these reasons, a molecule may not emit at the same wavelength in isolation as it did when paired with other fluorophores. For example, the molecule $\beta$-carboline has been isolated from fluorescent scorpions although its emission does not match the green color observed in the living organism\textsuperscript{[11]}. Likewise, a coumarin identified in scorpion sample jars after years of storage cannot account for the primary green emission of scorpions\textsuperscript{[12]}. Furthermore, all fluorophores are prone to photobleaching and in vivo mechanisms of protecting or replacing a fluorophore are lost after isolation.

Parallel factor (PARAFAC) analysis is one method of applying principal component analysis (PCA) to multi-dimensional data sets and is well suited to processing fluorescence excitation emission matrix (EEM) data. PARAFAC has been used to obtain fluorescence fingerprints from samples where the fluorophore composition is unknown. For example, PARAFAC has been used to identify the origin of estuarine water samples\textsuperscript{[13]} and to verify the exchange of ballast water from ships\textsuperscript{[14]}.

Here, we provide a comprehensive characterization of the fluorescent molecules present in a spider’s circulatory fluid (hemolymph) and show that spiders contain a complex mixture of fluorophores. By studying hemolymph extracts from multiple individuals using EEM fluorescence spectroscopy we were able to characterize the optical properties of their fluorophores without having to isolate them. Instead, the spectra of the individual components to fluorophores found in scorpions is shown that these spiders possess different fluorophores than scorpions. Furthermore, the fluorescence observed is not consistent with $\beta$-carboline or 4-methyl-$\beta$-hydroxycoumarin, two compounds previously described in scorpions.

Fluorophores from the hemolymph of yellow sac spiders (\textit{Cheiracanthium mildei}) have been characterized using excitation emission matrix (EEM) fluorescence spectroscopy. This approach provides characterization of fluorophores present in the organism without having to isolate pure samples. Minimal variation occurs between individual samples and each EEM has two distinct peaks, suggesting two fluorophores may be present in the hemolymph. Parallel factor analysis reveals that three fluorophores (with excitation and emission maxima at 270/319, 330/389, and 350/465 nm) best explain the sample to sample variation. By comparing the spectra of the three individual components to fluorophores found in scorpions it is shown that these spiders possess different fluorophores than scorpions.
## 2. Experimental

### 2.1. Data collection

Seven adult *Cheiracanthium mildei* spiders were collected in Portland Oregon and immediately frozen at −80°C. Their abdomens were removed and ground in 95% ethanol in a 1.5 mL tube using a disposable pestle. A *Uroctonus mordax* scorpion was collected from eastern Washington, frozen at −80°C, and the hemolymph from its abdomen was ground in 95% ethanol. Extractions were allowed to sit in the dark for 48 h at room temperature. Extractions were centrifuged at 14,000 rpm for 10 min. The supernatants from each of the seven samples (S1–S7) were transferred to a 1 cm quartz cuvette and analyzed on a steady state PTI QuantaMaster (Monmouth, NJ, USA) using 75 W Xe arc lamp excitation. 9H-pyrido [3,4-b]indole ([β-carboline] was purchased from Sigma–Aldrich and was dissolved in 0.1 N H2SO4 immediately prior to collecting an EEM.

EEMs were formed from emission spectra collected using excitation wavelengths from 250 nm through 475 nm in 5 nm increments. Each emission spectrum was scanned in 1 nm increments, starting from 10 nm above the excitation wavelength to a maximum wavelength of 650 nm. The signal at each emission wavelength was averaged for 0.5 s. Excitation and emission slit widths were set to a bandpass of 2 nm. The sequence of emission scans was alternated between high excitation and low excitation to eliminate the possibility that photobleaching would introduce systematic decreases in emission intensity. First-order Rayleigh scattering was avoided by starting data collection 10 nm above the excitation wavelength and Raman scattering was minimized by using 95% ethanol as a solvent. Excitation spectra were obtained by using a fixed emission wavelength and an excitation and emission bandpass of 2 nm. Absorption spectra in ethanol were collected on each sample using an Ocean Optics USB 2000 (Dunedin, FL).

### 2.2. Data processing

Primary and secondary inner filter effects were eliminated from the EEMs using a MATLAB routine [13] that applies a correction derived by MacDonald [15]. Individual absorption spectra for each sample were used to correct each EEM. Second-order Rayleigh scattering was removed from the EEM data by replacing with zeros any emission data that was greater than double the excitation intensity. First-order Rayleigh scattering was avoided by starting data collection 10 nm above the excitation wavelength and Raman scattering was minimized by using 95% ethanol as a solvent. Excitation spectra were obtained by using a fixed emission wavelength and an excitation and emission bandpass of 2 nm. Absorption spectra in ethanol were collected on each sample using an Ocean Optics USB 2000 (Dunedin, FL).

### 3. Calculation

Parallel factor (PARAFAC) analysis [17,18] and Tucker3 [19] are two fitting methods that apply PCA to multi-dimensional data. PARAFAC determines the combination of components that best describes a set of data with minimal residuals ($e_{ijk}$) using the model:

$$x_{ijk} = \sum_{j=1}^{f} a_{ij} b_{ij} c_{ij} + e_{ijk}$$

Each component is trilinear, consisting of two loading vectors (excitation and emission spectra) and one score vector (a weighting). To successfully fit EEMs with PARAFAC, the data must be trilinear and the amount of fluorophores must vary between samples. PARAFAC has been described as mathematical chromatography because it allows relative concentrations and spectra of individual components to be extracted from a group of EEMs. PARAFAC has been used to identify food impurities, [18] such as chlorophyll and hematoporphyrin in dairy products [20]. In a trial PARAFAC fit, the total number of components is varied and the results are analyzed to determine the number of components that provide the best description of the cube of data without using unnecessary components. Any increase in the number of components used will decrease residuals of the fit, however, if too many components are used, that reduction in residuals results from fitting noise rather than providing new information about the samples.

PARAFAC and Tucker3 analyses were performed using Matlab 7.4.0 (The Mathworks Inc., Natick, MA) and the N-way toolbox version 3.1 (http://www.maca.sourceforge.net/nwaytoolbox/index.asp) [17]. In all cases, the algorithms were constrained to nonnegative loadings and scores in all three modes and run to a stop criterion corresponding to a change in the sum-squared residuals (of the data to the model) of less than 1 × 10−6. Initiation was typically performed using single value decomposition after testing a number of different initiation methods, including random orthogonalization. Fits converged in less than 100 iterations. A core consistency diagnostic (N-way toolbox 3.1) was run for each trial fit. The core consistency [21] is a percentage of variation between a PARAFAC generated model and a model generated by Tucker3 [19]. Tucker3 is a data fitting methodology that fits multi-way data sets with the constraint of having orthogonal loadings. If the two programs generate identical components, as occurs when fitting to a single component, the core consistency will be 100%. Consistencies below 70% indicate that the two programs are producing different models; an indication that noise is being fit rather than data (over-fitting) [21]. Analyses were performed on a 1.83 GHz Intel Core Duo Macintosh.

A data set was assembled from all seven spider EEMs (S1–S7). PARAFAC fits were attempted using one through seven components. Each fit was repeated ten times and evaluated by comparing the sum-squared error, the appearance of residuals (intensity left out of the fit), by assessing the reasonableness and repeatability of the fits, and by performing the core consistency diagnostic. Fits to a single component were poor as evidenced by unreasonable peak appearance, high sum-squared error, and high residuals. Fits to two components resulted in high sum-squared error, structured residuals, and peaks that were asymmetric with $\lambda_{\text{max}}$ that varied when the fit was repeated. Fits to three components had low sum-squared error, minimal residuals, reasonable peak shapes, and were repeatable over dozens of trials. The average core consistency (Table 1) was calculated by repeating both the fits and the core consistency diagnostic ten times. The high core consistency of the three component fits indicates that each of the three components were necessary for fitting the data. In contrast, fits to more than three components gave inconsistent results, duplicative components, and low core consistency evaluations indicating over-fitting of the data (Table 1).

### 4. Results and discussion

#### 4.1. Excitation emission matrices (EEMs) of spider hemolymph

Fluorescence excitation and emission spectra collected on extracts from spider hemolymph reveal complex spectra. The sample (S1) shown in Fig. 1 is typical of the extracts studied. It revealed broad emission intensity with maxima near 390 and 450 nm when the sample was excited at 340 nm (Fig. 1A, solid line). In contrast, the excitation spectrum obtained from observation at 450 nm was relatively narrow and simple in structure (Fig. 1A, dashed line).

By collecting emission spectra at many excitation wavelengths, an EEM was produced that revealed additional information about...
the samples. A series of spectra, displayed as a contour plot (Fig. 1B) revealed two strong emission peaks centered near 390 and 450 nm with minor emission at higher and lower wavelengths. Each horizontal slice through the contour plot represents a single emission spectrum and likewise each vertical slice represents a single excitation spectrum. It is not possible to determine by inspection whether the fluorescence arises from a single fluorophore with a complex spectrum or from multiple fluorophores. Fig. 2 shows EEMs from extracts of six other *Cheiracanthium mildei* specimens (S2–S7) that were examined. The general contour was the same for each sample and the positions of the two primary emission peaks remained centered near 390 and 450 nm, however, the relative intensity of the two peaks varied. The variation in intensity between these two peaks suggests two different fluorophores with

---

**Table 1**

Core consistencies (CC) measured for 1 through 7 components (N)

<table>
<thead>
<tr>
<th>N</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100±0.0%</td>
</tr>
<tr>
<td>2</td>
<td>99.4±0.002%</td>
</tr>
<tr>
<td>3</td>
<td>91.6±0.16%</td>
</tr>
<tr>
<td>4</td>
<td>74.4±0.06%</td>
</tr>
<tr>
<td>5</td>
<td>18.0±6.94%</td>
</tr>
<tr>
<td>6</td>
<td>5.5±3.82%</td>
</tr>
<tr>
<td>7</td>
<td>1.09±0.03%</td>
</tr>
</tbody>
</table>

---

**Fig. 1.** Two different representations of fluorescence observed from an ethanol extraction of the spider *Cheiracanthium mildei*. (A) Single emission spectrum (solid line, excited at 340 nm) and excitation spectrum (dashed line, observed at 450 nm). (B) Contour plot representation of EEM of sample S1. Rayleigh scattering has been eliminated and inner filter effects have been corrected. Each isoline represents 10% of the sample intensity.

**Fig. 2.** Fluorescence EEMs of *Cheiracanthium mildei* samples S2 through S7. Each isoline represents 10% of the sample intensity. Rayleigh scattering has been removed and samples are corrected for inner filter effects.
varying amounts in each sample. This variation enables the use of parallel factor analysis to determine the individual fluorescent components.

4.2. Parallel factor analysis of fluorescence excitation emission matrices (EEMs)

The calculated spectra that resulted from the three component fits reveal three distinct fluorophores, two of which overlap in visible regions of the spectrum (Fig. 3). In all trials, the three component fits provided similar results with the components invariant across multiple trials. There are two components that emit largely in the visible portion of the spectrum and one that emits primarily in the UV. The double peak evident in the crude spectra is retained in component 1 (330 nm excitation/389 nm emission with a shoulder to ~450 nm). Component 2 resembles the long wavelength portion of that double peak, shifted to slightly higher emission wavelength (350 nm excitation/465 nm emission). Component 3 (270 nm excitation/319 nm emission) resembles the fluorescence EEM of tryptophan recorded in water [22] and one we collected in 95% ethanol (not shown). The weightings (Fig. 4), also referred to as scores, are a measurement of the amount of the components contained in each of the seven samples. While the weightings cannot be related to exact stoichiometries due to unknown fluorescence efficiencies the weightings do provide information on the relative contribution of each fluorophore to the overall fluorescence. The amount of components 1 and 2 varied to a relatively small degree between samples. The 91.6% core consistency indicates that three components are necessary to describe the samples. Fits with more than three components have lower core consistency (Table 1) indicating that the additional components used were not necessary.

Despite the appearance of two peaks in the spectra, the fit demonstrates that three distinct components best explain the variability. The whole organism extracts used here capture many sources of variability, including organism age and size that could limit the ability to fit data accurately. Samples that vary in the concentration of fluorophores will remain trilinear. However, changes in excitation and emission wavelengths would lead to non-trilinear data. The resultant model would be difficult to fit, and the core consistency of the fits would be low. The high quality of the fit obtained for these samples is an indication that the composition of fluorophores is invariant.

In contrast to our expectations, one fluorophore (component 1) is responsible for broad emission from 390 nm to 450 nm rather than two separate fluorophores at 390 and 450 nm. Since emission spectra from individual fluorophores usually mirror their excitation spectra (Kasha’s Rule), [23] a single fluorescent molecule seemed an unlikely explanation. However, the fit indicates that some of the intensity is best described by a single component with two discrete peaks at 390 and 450 nm. One possibility is that component 1 represents one molecule such as a protein or peptide that contains two fluorophores. If two fluorophores are linked in a single molecule their concentrations will remain fixed and they will behave as a single component. Another possibility is that an excimer is formed from two fluorophores resulting in the single excitation peak and two emission peaks [24].

While it is not possible to unambiguously identify contributing fluorophores from the data, the spectra of the components can be compared to known fluorophores to exclude candidate structures. For example, common biological fluorophores, such as NADH that fluoresces in ethanol 23 nm lower in wavelength than component 2 can be ruled out [25]. Furthermore, the β-carboline and coumarin previously identified in scorpions can be compared to each component. An EEM of ethanol-extracted scorpion hemolymph (Fig. 5a) and the molecule β-carboline (Fig. 5b) show no correspondence to any of the three components from the spiders. These types of comparisons are made possible by the collection and fitting of EEMs and would not be possible using simple, single emission

---

**Fig. 3.** Fit of Cheiracanthium samples (S1–S7) to three components using parallel factor analysis. (A) Emission spectra for components 1, 2, and 3 (arbitrary intensity). (B) EEM for components 1, 2, and 3.

**Fig. 4.** Weightings of component 1 (bottom, blue), component 2 (middle, green), and component 3 (top, red) in samples S1 through S7. The intensities are normalized to the full data set and the total bar heights reflect overall sample intensity for S1–S7.
Each isoline represents 10% of the sample intensity.

Fluorescence at each wavelength. As previously reported, spectra from the extracts where multiple compounds contribute to scorpion fluorophores. A: EEM from scorpion (U. mordax) ethanol extract. Each isoline represents 8% of the intensity. B: EEM of 9H-pyrido[3,4-b]indole (β-carboline). Each isoline represents 10% of the sample intensity.

Fig. 5. Scorpion fluorophores. A: EEM from scorpion (U. mordax) ethanol extract. Each isoline represents 8% of the intensity. B: EEM of 9H-pyrido[3,4-b]indole (β-carboline). Each isoline represents 10% of the sample intensity.

spectra from the extracts where multiple compounds contribute to fluorescence at each wavelength. As previously reported, β-carboline does not match the observed fluorescence in live scorpions and it is likely that additional fluorophores remain to be identified in scorpions. One possible explanation previously discussed is that β-carboline and coumarin are degradation products rather than the primary fluorophores [11].

The components identified show evidence of being biochemically related. Component 1 has spectral characteristics similar to component 2. The long wavelength portion of 1 resembles component 2. These observations are consistent with component 2 being a catabolic byproduct or precursor to the formation of 1. Furthermore, component 3, which emits at 319 nm is consistent with tryptophan which is a precursor to many fluorescent oxidation byproducts [26].

This method of analyzing fluorescence data can be useful for characterizing fluorophores from different species, sexes, or developmental stages of an organism. This will be useful to both chemists screening for natural products and to biologists assessing if there are differences in fluorophores among different organisms. It will allow for rapid discovery of whether different samples possess different fluorophores. In this way, it will allow the investigator interested in novel product discovery to focus their efforts on fluorophores with specific desirable optical properties.

5. Conclusions

Recent discoveries suggest that the occurrence of fluorophores in organisms may be widespread, however, only a few small molecule and protein fluorophore classes have been identified. Spiders are very diverse with more than 40,000 described species [27] and a rapid method of identifying fluorophores would allow screening of spiders from many divergent lineages. The small size of spiders makes it a challenge to extract, isolate, and identify individual fluorophores from a broad sampling of species. Benefits of the approach described here include speed, small sample size, the ability to study impure samples, and the avoidance of decomposition and photobleaching that can occur during separation and isolation.

This technique can accelerate the search for novel fluorescent proteins and small molecules that may be used in biotechnological applications. Fluorescent proteins [8] including green fluorescent protein (GFP), have become an important tool of biotechnology with applications in medicine, [28] sensors, [29,30] and imaging [31]. The screening technique described here will facilitate a search for previously unidentified fluorophores and a focus on fluorophores with specific properties that make them suitable for technological applications. Rather than screening isolated compounds, this technique uses crude hemolymph extracts in which the in vivo properties of the fluorophores are better maintained.

Acknowledgement

We gratefully acknowledge Michelle Knowles for assistance with Matlab.

References
