Identification and classification of silks using infrared spectroscopy

Maxime Boulet-Audet,¹,² Fritz Vollrath,² Chris Holland³

¹Department of Life Sciences, Imperial College London, London, UK
²Department of Zoology, University of Oxford, Oxford, UK
³Department of Materials Science and Engineering, University of Sheffield, Sheffield, UK

KEYWORDS

Infrared spectroscopy, cocoon, silk, multivariate analysis, phylogenetic
ABSTRACT

Lepidopteran silks number in the thousands and display a vast diversity of structures, properties, and industrial potential. To map this remarkable biochemical diversity, we present an identification and screening method based on the infrared spectra of native silk feedstock and cocoons. Multivariate analysis of over 1214 infrared spectra obtained from 35 species allowed us to group silks into distinct hierarchies and a classification that agrees well with current phylogenetic data and taxonomies. This approach also provides information on the relative contents of sericin, calcium oxalate, phenolic compounds, poly-alanine and poly(alanine-glycine) β-sheets. It emerged that the domesticated mulberry silk-moth *Bombyx mori* represents an outlier compared to other silk moth taxa in terms of spectral properties. Interestingly, *Epiphora bauhiniae* was found to contain the highest amount of β-sheet reported to date for any wild silk-moth. We conclude our approach provides a new route to determine cocoon chemical composition and in turn a novel, biological as well as material, classification of silks.
INTRODUCTION

Silkworm silk is a high-value agricultural product with growing technical applications (Borkner et al., 2014; Omenetto and Kaplan, 2010) and with sustainable harvesting contributing to socio-economic growth and poverty alleviation in rural communities (Astudillo et al., 2014; Dooley, 2004). Developments in mulberry sericulture and the increasing use of fiber from ‘wild’ silk worms provides the backdrop for increase interests in understanding the diversity of all silks. Not surprisingly, millions of years of divergent evolution have resulted in a rich biodiversity of silks (Scoble, 1999). Typically used in cocoons, this class of materials consists of a silk fibroin protein thread of up to one kilometer long coated with sericin proteins acting as resin/matrix glue (Chen et al., 2010b). This non-woven composite structure (Chen et al., 2010a) regulates gas flow and humidity (Danks, 2004; Horrocks et al., 2013; Roy et al., 2012), as well as protects the encased pupae from predation (Ishii et al., 1984), micro-organisms (Franceschi and Nakata, 2005) and the environment (Chen et al., 2012b). Silkworms produce cocoons with a broad variety of morphologies and architectures, ranging in porosity from loose meshes to full shells, with or without an exit opening (Chen et al., 2012c). Cocoons may also incorporate extraneous materials as well, such as integrated leaves for camouflage or an internally applied calcium oxalate solution that hardens the cocoon and may impart toxicity (Arnott and Webb, 2000; Chen et al., 2012c; Franceschi and Nakata, 2005; Gheysens et al., 2011; Takahashi Sy et al., 1969; Teigler and Arnott, 1972b). The diversity of lepidopteran silk materials includes the molecular dimension which amino acid analysis performed in the middle of the last century showing widely varying chemical compositions of silkworm silk (Hwang et al., 2001). However, while more advanced biochemical methods can inform on protein size (Hwang et al., 2001; Inoue et al., 2000; Mita et al., 1994), amino acid residue patterns (Navarro et al., 2008) and propensity to fold (Dicko et al., 2008), they are often labor intensive and
expensive. Hence, only a handful of fibroin proteins have been sequenced to date (Tanaka and Mizuno, 2001). Furthermore, these methods are focused on one specific molecular component of the cocoon and are unable to account for the other compounds present.

An alternative approach to achieve a broader chemical diversity assessment is to employ complementary spectroscopic and scattering techniques (Gheysens et al., 2011; Warwicker, 1954). For example, the use of attenuated total reflection infrared spectroscopy (ATR-IR) is particularly well suited to studying silks in all forms as it is capable of measuring rough and deformable solids (Chen et al., 2012a; Gheysens et al., 2011), as well as turbid and concentrated protein solutions (Boulet-Audet et al., 2011). Requiring only minimal sample preparation, ATR-IR can selectively probe the inside or outside surface of silk cocoons, providing information on the local chemical composition (Boulet-Audet et al., 2013; Chen et al., 2012a; Chen et al., 2007). After all, this spectroscopic method can determine (I) the level of protein crystallinity (Boulet-Audet et al., 2008), (II) secondary structure (Goormaghtigh et al., 2006), and (III) specific protein components such as sericin (Barth, 2000b; Teramoto and Miyazawa, 2003). Infrared spectra are also indicative of (IV) non-protein molecules present in silk such as amount of water (Boulet-Audet et al., 2011), calcium oxalate (Chen et al., 2012b; Gheysens et al., 2011) and carbohydrate (Lu et al., 2011; Schulz and Baranska, 2007).

In addition, the multivariate analysis of infrared spectra can (V) discriminate and classify samples based on their degree of relatedness. This infrared based classification approach can even discriminate bacterial species (Kansiz et al., 1999; Preisner et al., 2007), types of human hairs (Panayiotou and Kokot, 1999), and coffee bean varieties (Briandet et al., 1996) as well as providing information to construct taxonomic trees (Zhao et al., 2006).

In this study, we analysed unspun native silk feedstock from six species across the Saturniini and Attacini tribes also to spun silks from 35 species across the Lepidoptera and Arachnida. Multivariate and hierarchical clustering analysis performed on over 1000
individual spectra allowed us to build taxonomic trees and compared them to trees based on protein-coding nuclear genes (Chen et al., 2012c; Regier, 2005; Regier et al., 2008a; Regier et al., 1998; Regier et al., 2008b; Regier et al., 2002). As we shall demonstrate, we identified several interesting outlier species that produce silk with very different chemical compositions as well as provided a hypothesis as to their origins. Such species should have greater potential for use in industrial and biomedical applications than commonly employed silk.

RESULTS

NATIVE FEEDSTOCK SPECTRAL FEATURES: To evaluate the chemical variability of unspun silk without exogenous material, we used infrared spectroscopy to compare the native feedstocks of key species from the Lepidoptera genera *Actias, Attacus, Bombyx* and *Saturnia* and the spider *Nephila edulis* as the outgroup. Bombycidae feedstocks, such as *Bombyx mori* silk comprises of heavy and light chain fibroins as well as P25 linked with disulfide bonds (Chevillard et al., 1986; Mita et al., 1994) in a 6:6:1 ratio (Inoue et al., 2000). In contrast, feedstocks of the Saturniidae such as *Antheraea yamamai, Actias luna, Attacus atlas* and *Saturnia pavonia* (Tanaka and Mizuno, 2001) comprise a homodimer (double heavy chain, H-H) protein mixture. As arthropods, the spiders are very distantly related to the silkworms, yet by all accounts have evolved silk production independently around 400 million years ago (Craig, 1997). The similar flow properties of their feedstocks thus represent an excellent example of convergent evolution (Craig, 1997; Holland et al., 2006).
Figure 1 Infrared spectra of unspun native silk feedstock from domesticated silkworm silk (*Bombyx mori*), wild silkworm silk (*Attacus atlas, Antheraea yamamai, Actias luna, Saturnia pavonia*) and spider silk feedstock (*Nephila edulis* major ampullate). The 1700 to 1500 cm$^{-1}$ region is not shown as little differences between species were observed. Infrared spectra were collected from feedstocks extracted directly from the animal and kept at a native concentration (~22% DW).

Figure 1 illustrates the distinctive features of native silk feedstock infrared spectra between 900 and 1500 cm$^{-1}$ for silk from a variety of species. Table 1 indexes band assignments. Peaks between 1340 and 1456 cm$^{-1}$ are commonly assigned to the vibration mode of residues (Barth, 2000b). The strong 1383 cm$^{-1}$ band associated with CH$_2$ bending for wild silks (the top four curves in Figure 1) suggests a higher proportion of long-chain residues in feedstocks.
from *Bombyx mori* and *Nephila edulis* major ampullate silk glands. Another important distinction for wild silk feedstocks is the presence of the well resolved 1308 cm\(^{-1}\) peak in the amide III region. Monitored by Rheo-IR (Boulet-Audet et al., 2013), this band vanishes under shear-induced denaturation (see Figure S7) and is absent from cocoon spectra, and we have assigned this component to β-turns precursors to β-sheets formed after spinning (Bandekar and Krimm, 1980; Cai and Singh, 2004; Rousseau et al., 2006). Arginine-glycine-aspartic acid (RGD) residue pattern (Sukopp et al., 2002) are believed to procure greater fibroblast proliferation rate to the wild silkworm *Antheraea mylitta*’s compared to domesticated *Bombyx mori* silk (Minoura et al., 1995; Navarro et al., 2008). Hence, we hypothesised that RDG patterns might contribute to the wild silk specific vibration mode at 1308 cm\(^{-1}\). The amide III shoulder at 1270 cm\(^{-1}\) results from α-helices (Cai and Singh, 2004; Krimm and Bandekar, 1986), which also appears stronger in wild silkworm silk feedstock. The neighboring peak at 1245 cm\(^{-1}\), is commonly assigned to random coil secondary structures (Cai and Singh, 2004; Shao et al., 2005; Taddei and Monti, 2005; Yoshimizu and Asakura, 1990), strongest non-wild silks *Bombyx mori* and *Nephila edulis*. While present for all silk feedstocks, the peak at 1165 cm\(^{-1}\) associated with the stretching of the N-C\(_\alpha\), is clearly broader for non-wild species suggesting a wider distribution of conformations.

*Actias luna* is the only species probed showing a well-resolved peak at 1144 cm\(^{-1}\). We speculatively assigned this distinct band to the C-O stretching of sericin-like components used as binding resin/matrix as this species produces a cocoon with low porosity and high density (Chen et al., 2012c). While this study focused on unprocessed silk, extracting the sericin for further analysis could help clarifying this speculative assignment.

The 1103 cm\(^{-1}\) band appeared on all spectra collected, although was stronger on wild silkworm feedstocks. In the skeletal vibration region, this peak is likely to be caused by the C-C stretching of tyrosine aromatic rings, tryptophan or phenolic compounds (Andrus, 2006;
Barth, 2000b; Taddei and Monti, 2005). The adjacent component at 1075 cm\(^{-1}\) is present in all silk feedstocks (except sericin free spider silk) and is also observed in pure sericin spectra, but is strongest in *Actias luna* thus reinforcing our previous assignment of the 1144 cm\(^{-1}\) peak for this species (Anghileri et al., 2007; Barth, 2000b; Gupta et al., 1997). We also assigned the band at 1052 cm\(^{-1}\) to sericin C-O stretching, which is well resolved in most silkworm silks (Gupta et al., 1997; Taddei and Monti, 2005; Teramoto and Miyazawa, 2003). Table 1 Assignment of the main bands present in silk between 1700 and 1315 cm\(^{-1}\). (δ is for bending, \(\nu\) is for stretching, \(w\) is for wagging, \(r\) is for rocking, * is for “strongest observed in”)

<table>
<thead>
<tr>
<th>Position (cm(^{-1}))</th>
<th>Assignment</th>
<th>Observation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1733 (\nu(C=O))O</td>
<td>Tanned cocoon silks <em>O. eucalypti</em> <em>A. edwardsii</em></td>
<td>(Silverstein, 1981)</td>
<td></td>
</tr>
<tr>
<td>1699</td>
<td>Amide I, β-sheets/ β-turns</td>
<td>All spun silks <em>E. bauhiniae</em></td>
<td>(Bandekar and Krimm, 1979; Garside et al., 2005; Miyazawa and Blout, 1961; Moore and Krimm, 1976; Teramoto and Miyazawa, 2005)</td>
</tr>
<tr>
<td>1642</td>
<td>Amide I, unordered</td>
<td>All silks</td>
<td>(Boulet-Audet et al., 2008; Goormaghtigh et al., 2006; Jeong et al., 2006; Venyaminov and Kalnin, 1990)</td>
</tr>
<tr>
<td>1620</td>
<td>Amide I, β-sheets</td>
<td>All spun silks <em>E.bauhiniae</em></td>
<td>(Boulet-Audet et al., 2008; Moore and Krimm, 1976; Sonoyama and Nakano, 2000)</td>
</tr>
<tr>
<td>1547</td>
<td>Amide II, unordered</td>
<td>All silks</td>
<td>(Boulet-Audet et al., 2008; Goormaghtigh et al., 2006; Jeong et al., 2006; Venyaminov and Kalnin, 1990)</td>
</tr>
<tr>
<td>1516</td>
<td>Tyr–OH</td>
<td>All <em>A.yamamai</em> silks</td>
<td>(Chirgadze et al., 1975)</td>
</tr>
<tr>
<td>Position (cm⁻¹)</td>
<td>Assignment</td>
<td>Observation</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1508</td>
<td>Amide II, β-sheets</td>
<td>All spun silks *E. bauhiniae, A. panda</td>
<td>(Boulet-Audet et al., 2008; Moore and Krimm, 1976; Sonoyama and Nakano, 2000)</td>
</tr>
<tr>
<td>1456</td>
<td>δ_{as}(CH₃) Ala, Val</td>
<td>All unspun silks *N. clavipes major</td>
<td>(Barth, 2000b; Boulet-Audet et al., 2008; Colthup, 1964)</td>
</tr>
<tr>
<td>1443</td>
<td>δ_{as}(CH₃), β-sheets, (AG)n, (A)n</td>
<td>All spun silks *N. clavipes A. Panda</td>
<td>(Barth, 2000a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1417</td>
<td>δ₈(CH₃) Ala, Val</td>
<td>All unspun silks *N. clavipes major</td>
<td>(Barth, 2000a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1403</td>
<td>δ₈(CH₃) Ala, Val</td>
<td>All spun silks *E. bauhiniae</td>
<td>(Barth, 2000a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1395</td>
<td>δ(CH₂, OH) Serine</td>
<td>Silkworm silk *Outermost layer</td>
<td>(Anghileri et al., 2007; Barth, 2000b; Teramoto and Miyazawa, 2003)</td>
</tr>
<tr>
<td>1383</td>
<td>δ(CH₂) (AG)n</td>
<td>All unspun silks *A. attacus *S. pavonia</td>
<td>(Barth, 2000a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1370</td>
<td>δ(CH₂) (AG)n</td>
<td>All spun silks *E. bauhiniae</td>
<td>(Barth, 2000a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1340</td>
<td>δ(CH) or w(CH₂)</td>
<td>All unspun silks</td>
<td>(Barth, 2000b; Colthup, 1964)</td>
</tr>
<tr>
<td>1315</td>
<td>υ₅(OCO⁻) Calcium oxalate</td>
<td>G. postica outer cocoon Antheraea outer cocoon</td>
<td>(Chen et al., 2012c; Gheysens et al., 2011; Sargut et al., 2010; Silverstein, 1981)</td>
</tr>
</tbody>
</table>

Table 2 Assignment of the main bands present in silk between 1308 and 779 cm⁻¹. (δ is bending, υ is stretching)
<table>
<thead>
<tr>
<th>Page No.</th>
<th>Terms</th>
<th>Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1270</td>
<td>Amide III, α-helices</td>
<td>All silks</td>
<td>(Cai and Singh, 2004; Krimm and Bandekar, 1986)</td>
</tr>
<tr>
<td>1237-45</td>
<td>Amide III, random coil</td>
<td>All silks</td>
<td>(Cai and Singh, 2004; Shao et al., 2005; Taddei and Monti, 2005; Yoshimizu and Asakura, 1990)</td>
</tr>
<tr>
<td>1217</td>
<td>Amide III, β-sheets</td>
<td>All spun silks *E. bauhiniae</td>
<td>(Cai and Singh, 2004; Moore and Krimm, 1976; Shao et al., 2005; Taddei and Monti, 2005; Yoshimizu and Asakura, 1990)</td>
</tr>
<tr>
<td>1165</td>
<td>$\nu$ NCα</td>
<td>All silks *E. bauhiniae</td>
<td>(Barth, 2000b; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>1130</td>
<td>CH2OH Polyphenols</td>
<td>Tanned cocoon silks *O. eucalypti *A. edwardsii</td>
<td>(Lu et al., 2011; Schulz and Baranska, 2007)</td>
</tr>
<tr>
<td>1103</td>
<td>$\nu$(C-O), $\nu$(C-C), Polyphenols or tyrosine</td>
<td>Tanned cocoon silks Most wild silkworm silks *A. atlas</td>
<td>(Andrus, 2006; Barth, 2000b)</td>
</tr>
<tr>
<td>1068-75</td>
<td>$\nu$(C-O), $\nu$(N-C$_\alpha$) Serine</td>
<td>Silkworm cocoon *Outermost</td>
<td>(Anghileri et al., 2007; Barth, 2000b; Gupta et al., 1997)</td>
</tr>
<tr>
<td>1052-58</td>
<td>$\nu$(C$_\beta$-O) $\nu$(C-OH) Serine</td>
<td>Most silkworm silks *A. luna</td>
<td>(Gupta et al., 1997; Taddei and Monti, 2005; Teramoto and Miyazawa, 2003)</td>
</tr>
<tr>
<td>1028</td>
<td>$\tau$(CH$_2$), (A)$_n$, random coil</td>
<td>Unspun N. edulis and wild silks</td>
<td>(Moore and Krimm, 1976; Taddei et al., 2006)</td>
</tr>
<tr>
<td>1016</td>
<td>$\tau$(CH$_2$), (AG)$_n$, random coil</td>
<td>Unspun B. mori only</td>
<td>(Moore and Krimm, 1976; Taddei and Monti, 2005)</td>
</tr>
<tr>
<td>998</td>
<td>$\tau$(CH$_2$), (AG)$_n$, β-sheets *A. *panda *Bombyx</td>
<td></td>
<td>(Moore and Krimm, 1976; Taddei and Monti, 2005)</td>
</tr>
<tr>
<td>975</td>
<td>$\tau$(CH$_2$), (AG)$_n$, β-sheets A.panda and Bombyx</td>
<td></td>
<td>(Chen et al., 2012a; Moore and Krimm, 1976)</td>
</tr>
<tr>
<td>961</td>
<td>$\tau$(CH$_2$), Wild silks</td>
<td></td>
<td>(Moore and Krimm, 1976;</td>
</tr>
<tr>
<td>779</td>
<td>(A)_n β-sheets</td>
<td>Papadopoulos et al., 2007)</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>----------------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>δ(OCO') Calcium oxalate</td>
<td>(Chen et al., 2012c; Gheysens et al., 2011; Sargut et al., 2010; Silverstein, 1981)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*G. postica Antheraea outer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SILK COCOON SPECTRAL FEATURES: We have demonstrated that silk feedstocks have clear spectral differences between species. Therefore, we must assume that the cocoons produced from this feedstock should also show variability. Moreover, we would also expect this diversity to increase as construction introduces variables such as the larva’s spinning behavior, other silkworm secretions such as feces, and exogenous materials such as tannins diffusing from leaves. Previous work has shown that the properties of silk cocoons vary between the innermost and outmost layers (Chen et al., 2012a). Thus to examine these sources of chemical diversity, we compared the infrared spectra of the inner, and outermost layers of cocoons from 35 species of silkworms alongside the spectra of Nephila edulis dragline silk. Due to silk’s molecular alignment, spectra will vary depending on the fibres orientation relative to the beam path (Boulet-Audet et al., 2008; Papadopoulos et al., 2007). For a fair comparison against cocoon with no-preferential orientation (Chen et al., 2012b), spider silk filaments were arranged in a similar random orientation order.
Figure 2 a) Infrared spectra of key species innermost layer and *Bombyx mori* sericin spectrum for comparison. b) Outermost layer spectra from selected species.

Figure 2a shows spectra acquired from the innermost part of the cocoons from selected distinctive species. While the primary constituent of these cocoons is still made up of the silk proteins, the cocoons’ infrared signature differed substantially from these respective feedstocks. We assign these differences to a number of causes: the water content is lower in cocoons, reducing the ratio of amide I to amide II height (1642/1508 cm\(^{-1}\)), precursor helical
structures and random coils present in the feedstocks are converted via spinning into β-sheets, resulting in decreasing absorbance at 1642, 1547, 1308 and 1247 cm\(^{-1}\) and rising absorbance at 1699, 1620, 998 and 961 cm\(^{-1}\) (see Table 1 and 2). The relative absorbance of these β-sheet peaks can serve as an indicator of protein crystallinity. The peaks at 1699 and 1620 cm\(^{-1}\) in the amide I region are commonly used to determine the anti-parallel β-sheet content, but this overlaps with adjacent components from the fibroin and other compounds present in cocoons. In contrast, the low-frequency component at 961 cm\(^{-1}\) assigned to poly-alanine (A)\(_n\) is much better resolved (Moore and Krimm, 1976; Papadopoulos et al., 2007; Taddei and Monti, 2005). This band also appears in Nephila edulis dragline and most spectra of wild silks cocoons, particularly Epiphora bauhiniae (black). In contrast, some species like Bombyx mori (red) and Anaphe panda (magenta) have two weaker peaks at 998 and 975 cm\(^{-1}\) as their β-sheets are constituted instead of poly(alanine-glycine) segments (Moore and Krimm, 1976; Taddei and Monti, 2005).

Figure 2b demonstrates that the distinctive spectral features observed in the innermost layer of the cocoons were even more prominent in the outermost layer. For example, there was a higher relative absorbance of bands between 1395 and 1058 cm\(^{-1}\), which is consistent with the greater amount of sericin in the outermost layer for species like Bombyx mori. For comparison, a pure spectrum of sericin is included in Figure 2a (Chen et al., 2012a). Another clear difference between the two layers was the amount of calcium oxalate (Ca(COO)\(_2\)) crystals found embedded in the outermost layer of some species (Freddi et al., 1994; Freddi et al., 1993; Gheysens et al., 2011; Takahash.Sy et al., 1969; Teigler and Arnott, 1972b). Calcium oxalate vibration modes at 1315 and 779 cm\(^{-1}\) dominate the outermost layer of Gonometa postica cocoons yet were much weaker in the innermost layer. Another spectral contrast between layers was found in Opodipthera eucalypti, where the shoulder at 1733 cm\(^{-1}\)
can be assigned to the carboxylic acid and the polyphenol hydroxyls around 1000 cm$^{-1}$ (Andrus, 2006; Silverstein, 1981).

Thus, in summary, we established that it is possible to use structural and chemical markers to determine the type of crystallinity, the presence of sericin, calcium oxalate, and the polyphenol content in the measured cocoons.

COMPARISON OF SPECIES CHEMICAL COMPOSITION: For each spectrum collected, the interesting peaks identified in the previous section were integrated to estimate the relative content of calcium oxalate (1315 and 779 cm$^{-1}$), β-sheet crystallinity (1699, 1620, 1508, 998 and 961 cm$^{-1}$), the presence of tannin/phenolic compounds (1000 and 1733 cm$^{-1}$) and the sericin resin/matrix (1395 and 1058 cm$^{-1}$) (Figure 3).
Figure 3  

a) Relative area of a band assigned to calcium oxalate (740-800 cm\(^{-1}\)).  
b) Relative area of the band associated with \((A)_n \beta\)-sheets (931-983 cm\(^{-1}\)).  
c) Relative area of a \((AG)_n \beta\)-sheets assigned band (984-1006 cm\(^{-1}\)).  
d) Relative area of a band associated with tannins (1094-1135 cm\(^{-1}\)).  
e) Relative area of a sericin marker band (1384-1403 cm\(^{-1}\)).  

The outermost layer values are presented by brown bars, and the green bars shows the values of the innermost layer. The error bars represent the standard deviation of the different observations (n > 10). A value of 1 represent the highest area calculated and 0 the minimum measured.
CALCIUM OXALATE MINERAL CRYSTALS: Also called raphide, calcium oxalate forms highly toxic needle-like crystals, which can tear soft tissues and are thought to represent a plant defense mechanism (Arnott and Webb, 2000). Because no known metabolic pathways convert calcium oxalate in silkworms, we assume that calcium oxalate presence in the cocoon is a result of the ingestion of leaves containing the compound and resultant excretion by the silkworm. While this may be the case for wild silkworms, it appears that artificial selection has changed the behavior of *Bombyx mori* silkworm to prevent this excretion into the cocoon. Figure 3a shows the relative intensity of the band at 779 cm\(^{-1}\) achieved by integrating the absorbance between 740 and 800 cm\(^{-1}\). This well-resolved band was used as a relative indicator of the amount of microscopic calcium oxalate monohydrate (Ca(COO)\(_2\)) mineral crystals present on the cocoons (Chen et al., 2012b; Gheysens et al., 2011). Our ATR-IR results identified cocoons from *Gonometa postica* to have the highest calcium oxalate content. The host plant of *Gonometa postica*, *Acacia*, is also rich in calcium oxalate to detoxify calcium ions (Franceschi and Nakata, 2005; Martin et al., 2012; Teigler and Arnott, 1972a). The high calcium oxalate content of *G. postica* cocoons and *Antheraea* genus measured are in agreement with previous reports and electron microscopy observations on these cocoons (Chen et al., 2012b; Gheysens et al., 2011). Cocoons from *Samia*, *Hylophora* and *Attacus* also indicate the presence of calcium oxalate, but in lower proportions while other species measured have only minute amounts on their cocoon.

The presence of calcium oxalate in the cocoon is known to complicate the industrial reeling as it prevents long lengths of fiber extraction (Gheysens et al., 2011). Previous work has shown that while calcium oxalate is notoriously toxic to humans and responsible for kidney stone formation (Evan et al., 2007). The commonplace edetic acid (EDTA) treatment for dissolving kidney stones was found to be equally effective at demineralizing wild silk cocoons containing calcium oxalate crystals and enabling industrial processing (Gheysens et
al., 2011). Thus the ability to detect and quantify the amount of calcium oxalate present in
a cocoon prior to processing may have industrial advantages in minimizing reagent use or in
selecting low mineral content cocoons in the first place.

**Β-SHEET CRYSTALLINITY:** X-ray scattering initially proved the presence of β-sheet
nanocrystals inside silk fibers (Warwicker, 1954). Conveniently polyalanine (A)$_n$ and
polyalanine glycine (AG)$_n$ β-sheet structures also give distinctive peaks in silk infrared
spectra, indicative of the degree of crystallinity present and by extension may relate
mechanical properties (Boulet-Audet et al., 2008; Moore and Krimm, 1976; Porter et al.,
2005; Sonoyama and Nakano, 2000). Using the integrated absorbance of polyalanine
antiparallel β-sheets peaking 931 to 983 cm$^{-1}$, Figure 3b shows the relative (A)$_n$ β-sheet
content across the species tested. Our results suggest that *Epiphora bauhiniae* has the highest
degree of crystallinity amongst the all (A)$_n$ containing silks measured, followed by species
from the *Samia*, *Antheraea*, and *Attacus* genera. For most species, the (A)$_n$ β-sheet content
appears greater in the innermost layer, most likely due to non-fibroin compounds contributing
to the infrared signal more on the outermost layer. In comparison, spider silk dragline from
*Nephila edulis* appears to have a comparable (A)$_n$ β-sheet crystallinity compared to most
silkworm silks measured. On the lower end, *Gonometa*, *Argema*, and *Caligula* genera seemed
to have the lowest (A)$_n$ β-sheet content amongst all the species studied. Integrating the region
984 and 1006 cm$^{-1}$ quantified the contribution of the (AG)$_n$ peaks at 975 and 998 cm$^{-1}$ while
excluding the (A)$_n$ β-sheet peak at 961 cm$^{-1}$. Only three species appeared to have (AG)$_n$
repetitive segments, *Bombyx mori*, *Bombyx mandarina* and *Anaphe panda*. Unlike
Bombycidae and Noctuidae families, none of the Saturniidae cocoons displayed peaks
associated with the (AG)$_n$ structure (Moore and Krimm, 1976; Taddei and Monti, 2005). This
fundamental distinction could be related to their appurtenance to different taxonomic families
(see Figure 6).
TANNINS AND PHENOLIC COMPOUNDS: Wild silkworms naturally secrete some phenolic compounds in their silk (Brunet and Coles, 1974), but our results suggest that that additional hydroxyl-containing compounds, such as polyphenols, could come from exogenous sources. By integrating the absorbance between 1035 and 1094 cm\(^{-1}\), the relative amount of these molecules can be estimated. Figure 3c shows that a few species had phenolic compounds located mainly on the outside of the cocoon including *Opodipthera eucalypti*, *Saturnia pyri*, *Hyalophora gloveri*, *Attacus edwardsii*, *Antheraea polyphemus* and *Actias luna*. This finding agrees with the hypothesis that leaves incorporated by the silkworm into the cocoon structure leech water soluble plant polyphenols when wet. In contrast, species that do not integrate leaves into their cocoons such as *Antheraea mylitta* and *Attacus atlas* showed low phenolic compound parameter scores.

SERICIN PROTEIN GUM: Sericin proteins are essential to the cocoon construction as they are used to bond fibers together (Chen et al., 2012a). The amount of sericin present in a cocoon can be inferred from the absorption bands between 1384 and 1403 cm\(^{-1}\) associated with the amino acid serine, an amino acid present in high quantities in sericin but not in fibroin (Teramoto and Miyazawa, 2003). Figure 3e suggests that *Bombyx* genus silks have the most sericin along with *Actias*, *Antheraea*, *Saturnia* and *Samia* genera silks. Our results suggest that less sericin in the coats of the high porosity cocoons of species such as *Caligula trifenestrata*, *Graela isabella* and *Loepa katinka* (Chen et al., 2012b). Differences in sericin abundance between the innermost and outermost layers of the cocoons tested indicate that *Bombyx mori*’s cocoons have more sericin in the outermost layer, consistent with previous findings (Chen et al., 2012a). However due to the additional mineral and phenolic components of the wild silks, it is challenging to interpret the distribution reliably in the other silks tested.
SILK SPECIES CLASSIFICATION: Our results show that the integration of IR bands assigned to individual compounds can provide select windows into a silk cocoon’s chemical composition. However, single variable analysis exploits only a small fraction of the information contained within the spectra with thousands of data points. In contrast, multivariable analysis is far more powerful to classification and discriminate samples. Hence, we first performed a Principal Component Analysis (PCA) (Pearson, 1901) to reduce the number of variables while retaining most of the variability. The first principal component expresses the largest variance between samples. The PC scores, the relative importance of these PC for each spectrum, were subsequently used for the linear discrimination analysis (LDA) to model the differences between species with a set of factor coefficients and scores.

The LDA scored were able to discriminate broadly between wild and domesticated silks as well as spider silk. Of the 25 measurements selected randomly for validation, the method identified the correct species for 100% of the “unknown spectra” (see supplementary material). Figure S8 shows the tree generated from the LDA scores using hierarchical clustering analysis (HCA). However, as previously noted, once spun, even more variables are introduced into the cocoon structure, and thus our multivariate approach becomes even more powerful. The multivariate analysis had an identification hit rate of 70% for species and 75% for genus, tested using the randomly selected validation group of 200 “unknown spectra” (see supplementary materials).
Figure 4 First and second factor scores of the cocoon spectra contributing to 62% of species discrimination of the linear discriminant analysis.

Our initial multivariate analysis of cocoon diversity is summarized in Figure 4 which highlights the values of the first and second factor scores calculated from the cocoon spectra. The primary cluster encompasses most silks from wild silkworm species with *Antheraea* silks near its centroid (green markers). *Antherina suraka*, *Leopa katinka*, *Epiphora bauhiniae* and *Samia Cynthia* silks appear in cluster’s periphery, suggesting a greater dissimilarity with the average of the measured silks. Clearly discriminated species outside this cluster such as *Anaphe panda*, *Bombyx mori* and *Bombyx mandarina* appear as outliers. The spider *Nephila edulis* dragline silk is also outside the primary cluster, and easily discriminated from silkworm cocoons with the second factor scores. Noteworthily, our LDA implies that *Epiphora bauhiniae* is the silkworm species producing the closest silk to *Nephila*.
*edulis* dragline. However, more species from other families would need to be studied to attest to which of the thousands silkworm species spins ‘spider silk’. To develop this analysis further and begin to draw quantitative links between species, our HCA used the scores of the ten most important factors to group these species according to their similarity.

Figure 5 a) Cladogram generated from the phylogenetic analysis of Regier et al. (Chen et al., 2012c; Regier, 2005; Regier et al., 2008a; Regier et al., 2008b; Regier et al., 2002) b) Ultrametric tree generated from the hierarchical clustering analysis of cocoon infrared spectra LDA factor scores. Species with a Euclidean distance smaller than 525 were grouped together.

**GROUP 1:** *CALIGULA, SATURNIA, AND ACTIAS:* Illustrated in brown on Figure 5b, Group 1 encompasses *Caligula, Saturnia,* and *Actias* genus together along with *Opodipthera eucalypti* and *Cricula trifenestrata.* As most species from this group had high absorbance...
between 1094 and 1135 cm\(^{-1}\) (Figure 3d), this result suggests that these species were regrouped together partly based on their high phenolic content. Except *Cricula trifenestrata*, these species cocoons appeared substantially tanned with a dark brown coloration (Chen et al., 2012c). Also, these species do not present calcium oxalate crystals on their surface (Chen et al., 2012c), as confirmed in Figure 3a. Group 1 also appears to have lower \(\beta\)-sheet content than the other groups.

GROUP 2: *ARGEMA*: Group 2 contains the *Argema* genus. Unlike cocoons from group 1 they do not appear to have a high phenolic content or calcium oxalate and although largely comparable to neighboring groups these two factors could explain the large Euclidean distance with group 1.

GROUP 3: *ANTHERAEA*: From our classification, it appears that *Antheraea* silks all have small Euclidean distances relative to one another and as such they were all regrouped together with *Antherina suraka* in Group 3. Previous studies based on morphological feature classification argued that *Antherina suraka* could be more closely related to the African Bunaeni tribe than other species of the Saturniini (Oberprieler, 1997). The comparable Euclidean distance between *Antherina suraka* and *Antheraea frithi* weakens this hypothesis. Since *Antheraea* is the genus with the most calcium oxalate (Figure 3a), the linear discrimination analysis method could have regrouped these silks mainly on mineral content. However, *Antherina suraka* and *Antheraea frithi* show less absorption between 740 and 800 cm\(^{-1}\) (calcium oxalate) and are more distant from the other species of this group. In addition, Group 3 has weaker phenolic compound bands than Group 1 and yet has an average amount of sericin. The next closest species to these groups are *Loepa katinka* and *Graellsia isabellae* which both present low sericin content and high porosity according to other studies (Chen et al., 2012b; Chen et al., 2012c).
GROUP 4: ATTACUS AND SAMIA: Much more distant are the species classified in Group 4, including Samia, Hyalophora, and Attacus genera along with Callosamia promethea. When compared to previous reports, the morphology of the cocoons classified together in Group 4 (blue) appears characteristic as the innermost layers are much more compact than their outer layers (Chen et al., 2012c). This morphological difference could explain the higher amount of sericin measured on the innermost layer (Figure 3e). Adding to the composition variation between the inner and outmost layers these cocoons have an intermediate content of (A)_n β-sheets, sericin, and tannin when compared to those from Groups 1 and 3. Branching from Group 4, Epiphora bauhiniae has the largest amount of (A)_n β-sheets, little phenolic compounds, no calcium oxalate and little sericin.

GROUP 5: BOMBYX: The LDA placed Bombyx mori and Bombyx mandarina silks into a distant group. Even though mandarina appears to have more (AG)_n β-sheets than mori (Figure 3c), the difference between their spectra is subtle in comparison with the other species presented. This result suggests that the artificial selection of Bombyx mori might have played a lesser role than natural selection in differentiating this species from other Lepidoptera families.

SILK FROM OTHER SUPERFAMILIES: While still very distant, Anaphe panda had the smallest Euclidean distance to Bombyx. Our results suggest that Anaphe panda also has (AG)_n β-sheets, sericin and no calcium oxalate. Testing more silks from these two families would confirm whether these silks share the same spectral features. Social spinning behavior is another interesting characteristic of Anaphe panda, which partners with many other worms to build a communal cocoon nest (Mbahin et al., 2007). The cocoon’s structure does not depend on the silk quality of a single individual, resulting in different natural selection constraints. Also from another superfamily (Lasocampiadae), Gonometa postica silk is very distinct from all other species studied. The innermost layer appears closer to Saturniidae silks.
(Figure 2), whereas the major difference between their outer layer is likely to come from the large amount of calcium oxalate present.

COMPARISON OF ATR-IR AND PHYLOGENETIC TREES: Presented on Figure 5a, the ultrametric tree generated from the infrared spectra was compared to the phylogenetic tree built from the sequencing of a few protein-coding nuclear genes by Regier et al. (Chen et al., 2012c; Regier et al., 2008a; Regier et al., 2008b; Regier et al., 2002). The genes selected to construct this phylogeny produce proteins other than silk with various enzymatic functions such as carbamoylphosphate synthetase, aspartate transcarbamylase, dihydroorotase (Moulton and Wiegmann, 2004), dopa decarboxylase (Fang et al., 1997), enolase (Farrell et al., 2001) and wingless (Brower and DeSalle, 1998). Although quantitative a comparison between ultrametric and a unitless tree is not possible, they are strikingly similar, except for few species.
DISCUSSION

By assessing the diversity of wild silks, this study compared the biochemical composition of native silk feedstock from 6 species and silk cocoons from 35 species using infrared spectroscopy and multivariate analysis. For unspun native silk feedstocks, we identified a new spectral markers unique to wild silkworm silks, which we assigned to β-turn secondary structures. The hierarchical clustering of the feedstocks also profiled the dissimilarity of Saturniidae silks to the silks of Bombycidae and spiders.

Collecting spectra from silkworm cocoons provided information not only on the spun fibre but also on the non-protein chemical content and distribution across the layers. The Specific infrared bands revealed the relative content of sericin, calcium oxalate, phenolic compounds, polyalanine and poly(alanine-glycine) β-sheets. The multivariate analysis also permitted the hierarchical classification of 35 species (including one spider silk) into groups based on their chemical composition. This analysis revealed the presence of interesting outlier species with very dissimilar spectra, which could represent distinctive mechanical or chemical properties. Amongst these outliers were *Gonometa postica* cocoons, which had the highest calcium oxalate of all species measured. Furthermore the species with the most β-sheets, *Epiphora bauhiniae*, also appeared to have the closest chemical composition to *Nephila edulis* spider silk dragline. The *Bombyx* genus stood out from all other species measured, representing an outlier group. Consequently, using *Bombyx mori* as the model species for silk studies could lead to conclusions that are not applicable to all types of silks. Although our sampling had a bias towards Saturniidae silk, *Antheraea* silks were found to have median PC scores, suggesting that *Antheraea* silks are more representative of silk’s biodiversity. Not only did the multivariate analysis had a species identification hit rate of 70%, but the ultrametric trees were created from the infrared spectra.
Our analysis thus suggests a relationship between non-silk coding nuclear genes selected by Regier et al. and the silkworm cocoon’s overall biochemical composition. Such link implies that infrared spectra could be used as a proxy for species phylogenetic classification. Despite vast similarities between these trees, a few silk species were classified differently under these two approaches. This dissimilarity could be the result of non-protein based variation such as temperature or humidity or the incorporation of exogenous material onto the cocoons. For instance, *Cricula trifenestrata* was expected to be closer to the Antheraea silks rather than classified in Group 1. The difference could be due the fact that *Cricula trifenestrata* lives in an environment with a warm climate, requiring more ventilation than Antheraea silk cocoons found in colder regions (Kakati and Chutia, 2009). Interestingly, *Graellsia isabellae* silk should have been very similar to Actias silks, but is classified by our analysis outside Group 1 along with *Leopa katinka*. Their separate classification could result from the high concentration of tannins measured on the cocoons of these species. Oberprieler et al. suggested that these species might have been misclassified (Oberprieler and Nassig, 1994), and our study strengthens the hypothesis that Graellsia and Actias are two distinct genera. As expected, *Epiphora bauhiniae* was classified in the Attacini tribe but is rather distant from the other species of Group 4, most likely because of its higher (A)$_\beta$-sheet crystallinity content. In summary, despite minor differences in the classifications, our method represents a powerful but straightforward hierarchical classification tool to help resolving some of ambiguity in the relationships of Lepidoptera species.

Because of the intense selection pressure on this vital biological structure, we believe silk cocoons represent a model for the phylogenetic analysis of all silk-moths species. This untapped proxy method not only adds to more traditional gene and protein sequencing but is also less time-consuming and cheaper than e.g. protein sequencing. As silk cocoons are commonly part of entomology collections spanning hundreds of years of collecting, they can
be readily sourced and rapidly tested in a non-destructive manner. Such powerful longitudinal studies could shed light on silk-moth evolution and ecology by helping to resolve some of the relationships ambiguities of Lepidoptera species. Furthermore, with the advent of affordable handheld IR instruments our approach could also allow such analysis to take place in the field. Thus, combining ATR-IR with multivariate analysis could aid to unravel the evolution and biodiversity of silk producing species as well as inform us regarding which species is best suited to a particular industrial application.

METHODS

NATIVE SILK FEEDSTOCK PREPARATION: All wild silkworm eggs were purchased from Worldwide Butterflies (WWB, Dorset, UK). Actias luna, Antheraea yamamai, Attacus atlas, Saturnia pavonia were fed with Walnut (Juglans regia), Hawthorn (Crataegus monogyna), Privet (Ligustrum vulgare) and Hawthorn (Crataegus monogyna), respectively. When larvae started spinning their cocoon, native silk feedstocks were extracted from the silk glands of last instar silkworms. Final instar Bombyx mori worms were fed with white mulberry leaves (Morus alba). Nephila edulis major ampullate glands and dragline were extracted from mature female spiders fed with Drosophila spp. and Caliphora spp. and reared in-house under controlled temperature and humidity as described elsewhere (Holland et al., 2006).
Figure 6 Summary of higher-level relationships of the superfamily related to species studies adapted from Regier et al. (Regier et al., 2008a). The bracketed numbers represent the number of species measured in the superfamily. Adjacent pictures show the cocoons of species measured.

SILK COCOON PREPARATION: We analyzed cocoons from 35 different species belonging to the superfamily, Saturniini and Attacini, highlighted in Figure 6. icipe (African Insect Science for Food and Health) in Kenya provided *Gonometa postica* silkworm cocoons. The other cocoon species were purchased from WWB; the species chosen represented the four families of Lepidoptera. At least four 3.5 mm round cocoon disks were cut from each cocoon using a plier punch for analysis by IR spectroscopy.

SPECTRAL ACQUISITION AND TREATMENT: A Golden Gate single bounce diamond ATR accessory (Specac Ltd., London, UK) coupled to a Nicolet 6700 FTIR spectrometer equipped with a MCT nitrogen cooled detector (Thermo Scientific, Madison, WI) was used.
for spectra collection. Spectra acquisition was performed at a 4 cm⁻¹ resolution from 500 to 6000 cm⁻¹, averaging 32 to 64 scans at a 5.06 cm/s mirror speed. Although the fibers in a cocoon samples are randomly oriented, spectra were collected with the IR beam polarized perpendicularly to the plane of incidence (s) with a zinc selenide holographic wire grid polarizer (Thermo Scientific, Madison, WI). The ATR diamond’s internal reflection element (IRE) had a refractive index of 2.417 and an angle of incidence of 45°. For this configuration, the evanescent wave emerging out of the IRE could probe around 1.2 μm deep into the sample, the penetration varying with the wavelength (Harrick, 1967). The liquid state of native feedstock spectra ensured a good contact with the IRE for data collection. As the cocoons have an inherent roughness greater than tens of microns, an anvil was used to press on the cocoon disks to ensure a good contact with the IRE. For acquisition consistency, the pressure applied on cocoon disks was kept to the minimum necessary to obtain an absorbance of 0.1 for the amide II band. By aiming to keep the absolute absorbance consistent the anomalous dispersion of the refractive index were therefore comparable for each spectrum collected (Boulet-Audet et al., 2010). Before each measurement, the crystal was cleaned with tissues and demineralized water before a new background was acquired. This method helped compensating for the detector’s signal fluctuations as well as preventing contamination between measurements. The innermost and outermost layers of these disks were measured by collecting at least 18 distinct spectra from each species for a total of 1185 spectra across the 35 species studied. This study thus represents the spectroscopic study encompassing the largest number of wild silks type to date.

DATA PRE-PROCESSING: Spectral operations were performed using OMNIC 7.3 (Thermo Scientific, Madison, WI) using a custom VBA code. An offset was first subtracted to all spectra as calculated from the average of the region from 1950 to 1900 cm⁻¹. Spectra were then normalized using the integrated absorbance from 1900 to 800 cm⁻¹ to compensate for
absolute signal variations incurred by differing cocoon contact with the IRE. For the single component analysis, the relative area of each peak integrated was calculated by subtracting a linear baseline between the interval limits from the integrated absorbance.

MULTIVARIATE ANALYSIS AND DENDROGRAM GENERATION: A more detailed description of the multivariate analysis is presented in supplementary material. A principal component analysis (PCA) was performed on the first derivative spectra using Pearson’s method (Pearson, 1901). The ten first principal components (PC) were used for native silk feedstock and 40 first PCs for cocoon spectra. To discriminate between species based on infrared spectra, the dataset was randomly divided into the training and validation groups for the linear discriminant analysis (LDA) (Fisher, 1936; Yang et al., 2005). The Ward’s top-down hierarchical clustering method generated the phylogenetic tree dendrograms with the dissimilarity Euclidian distances (Mariey et al., 2001; Ward, 1963).
AUTHOR INFORMATION

Corresponding Author

*m.boulet-audet@imperial.ac.uk

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

European Research Council (SP2-GA-2008-233409), Magdalen College Oxford, the UK

EPSRC (EP/K005693/1; EP/G068224/1) and the Canadian NSERC (PGS 3D/6799-379132-2009)

Author competing interests

The authors declare no competing interests.
ACKNOWLEDGEMENTS

The authors wish to thank Dr Fujia Chen, Dr Ann Terry, Dr David Porter and Dr Beth Mortimer for his critical comments on the manuscript. The authors also wish to thank Dr Fujia Chen, Julia Van Campen, Alexander Greenhalgh and Addis Akebede for assisting with the data collection.

ABBREVIATIONS

FTIR, Fourier Transform Infrared; Linear Discriminent Analysis; PCA, Principal Component Analysis
REFERENCES


