**Instructions for R script: FTIRbaselines.R**

Suzanne B. Hodgkins, Ph.D

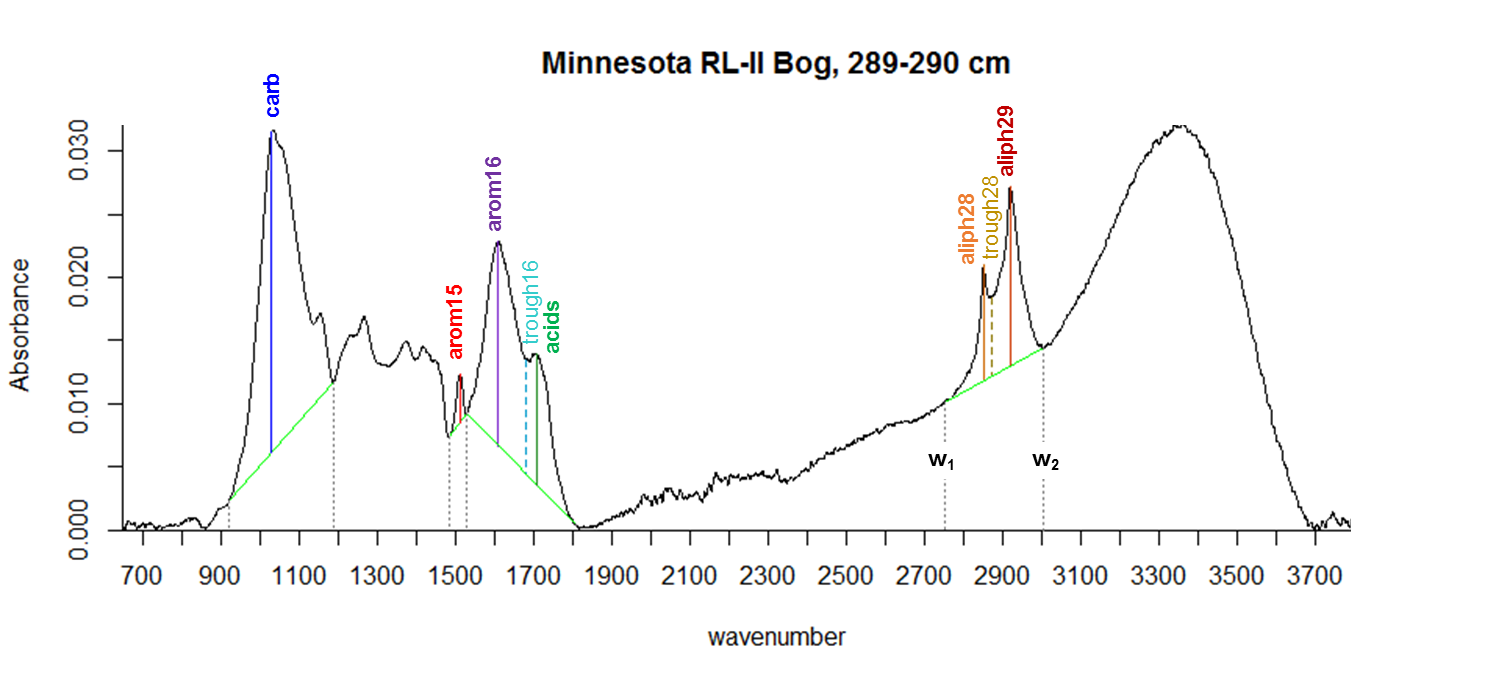
October 25, 2016

Overview

This is an R script for finding the location and absorbance of peaks (Table S1, Figure S1), including baseline-corrections, in FTIR spectra of plant and peat samples. It was written and tested with R version 3.2.1, running with RStudio version 0.99.467.

**Table S1.** Peaks and troughs found by the R script.

|  |  |  |
| --- | --- | --- |
| Peak Name | Approx. Wavenumber | Structural Assignment |
| *Peaks:* | | |
| carb | 1030 | carbohydrates |
| arom15 | 1510 | aromatics |
| arom16 | 1630 | aromatics, or deprotonated COO− |
| acids | 1720 | protonated COOH |
| aliph28 | 2850 | aliphatics |
| aliph29 | 2920 | aliphatics |
| *Troughs treated as peaks (i.e., have baseline corrections applied):* | | |
| trough16 | 1690 | trough between arom16 and acids |
| trough28 | 2870 | trough between aliph28 and aliph29 |



**Figure S1.** Example FTIR spectrum (Brunei peat), labeled with peaks (defined in Table S1) and baselines (bright green). w1 and w2 are labeled for the aliphatic baseline, and are defined similarly for other peak baselines.

Instructions

1. Save the FTIR spectra (in absorbance mode) to a .csv file with the wavenumbers in descending order in the first column. Make sure there are no extra rows except for the names of the samples in the header row.
2. Put a copy of the “FTIRbaselines.R” file in the same folder as the FTIR data.
3. Open and run the R script (in RStudio, click “Source”).
4. Follow the prompts (details as follows):
   1. Enter name of csv with FTIR absorbance data:  
      Enter the name of the file you created (including “.csv”, without quotes), and press enter. If the file is correctly formatted, it will then process the spectra while displaying the progress.
   2. Prompt for notes on each spectrum, which will be stored in output? (y/n)  
      **If you type ‘y’**, it will show each spectrum with baseline corrections drawn, and ask you to type something and press enter for each spectrum. Whatever you type will be stored in the Notes.csv output file (see Table S2) after the script is finished.  
      **If you type ‘n’**, it will still show the spectra, and you can look at them using the right and left arrows above the plots in RStudio, but you won’t be able to type any notes to be stored in the output file. This is faster if you have a lot of spectra and are confident that the baseline corrections will be accurate.  
      *In these plots, the true peaks are shown as solid lines while the troughs are shown as dashed lines, each with a unique color (Figure S1). When checking the spectra, make sure that all peaks, troughs, and baselines (bright green, with grey dotted lines connecting the endpoints to the x-axis) are in their expected locations.*
   3. Show spectra again, zoomed in on the region from 1450-1850 cm^-1?  
      **If you type ‘y’**, it will show the spectra again zoomed in on this region. This is a good idea if you have low or highly variable arom15, arom16, or acids peaks, as these peaks tend to be erroneously assigned more often than other peaks. Whatever preference you typed for “Prompt for notes on each spectrum…” also applies here; i.e. if you asked to type notes for the whole spectra, you will also get to type them here; otherwise it will just show them without prompting for notes.   
      **If you type ‘n’**, it will not show the zoomed-in region at all.
   4. ENTER NAME OF FOLDER FOR OUTPUTTING DATA (if different from original csv filename):  
      If you press enter without typing anything, it will save the output data to a folder with the same name as the original csv file. If you don’t want this, type what you want for the output folder name. Any special characters will be changed to underscores.
5. You may now either close R, or do further data analysis. (If doing more analysis in R, see the write.csv lines at the end of the script to see which variable names correspond to which dataset listed in Table S2; e.g. the variable Ap contains the data outputted to Raw.Peaks.csv.)
6. (optional) Combine the csv files into one Excel file.

**Table S2.** Data included in each output CSV file.

|  |  |
| --- | --- |
| CSV Filename | Data |
| Wavenumbers | Wavenumbers of all peaks and peak boundaries that are treated as peaks (Table S1). |
| W1 | Wavenumbers used for lower boundaries (w1) of each peak. |
| W2 | Wavenumbers used for upper boundaries (w2) of each peak. |
| success.W1 | For each peak, whether w1 could be found based on the spectra shape (value of TRUE), or a default value was used instead (value of FALSE). |
| success.W2 | For each peak, whether w2 could be found based on the spectra shape (value of TRUE), or a default value was used instead (value of FALSE). |
| Raw.Peaks | Raw absorbances for each peak (without baseline corrections). |
| Corr.Peaks | Baseline-corrected absorbances for each peak. |
| Norm.Raw.Peaks | Raw absorbances normalized to spectral area. |
| Norm.Corr.Peaks | Baseline-corrected absorbances normalized to spectral area. |
| Area.and.780 | Area of each spectrum. Also includes the absorbance at 780 cm−1, which represents silicate minerals. Both the raw (silicate780) and area-normalized (norm.silicate780) absorbances, equivalent to the “Raw.Peaks” and “Norm.Raw.Peaks” at this wavenumber (respectively). This peak is kept at a constant wavenumber and has no baseline correction applied; hence its inclusion in a separate file. If norm.silicate780>0.00005, the sample likely contains silicate minerals, which produce a large peak that interferes with the carb peak. |
| Notes | Type of aliphatic peaks (Table S3) and acids peak (Table S4). Also includes any other notes you type when looking at the spectrum graphs. |

Details on peak finding procedures

For each spectrum, the first and second derivatives are first calculated using linear regressions through a 29-point window, which I chose because it minimizes noise effects while still allowing the identification of broader spectral features.

The endpoints of each baseline (w1 and w2) are found by locating the local minimum within the expected region of each endpoint. If there is no local minimum, the maximum of the second derivative (i.e., point where the spectrum is the most concave-up) is used instead. If there are no concave-up points, w1 or w2 is set to a default value, and “success.W1” or “success.W2” for that peak is set to FALSE.

Once the baselines are assigned, baseline-corrected absorbances are calculated for each point above the baseline, and the peaks are then identified based on the maximum baseline-corrected absorbance. Thus, for slanting baselines, the peak locations may appear to be slightly offset.

Locations of peaks that share a common baseline (aliph28, trough28, and aliph29; and arom16, trough16, and acids) are determined slightly differently, depending on the degree of peak separation (which determines the peak “type”). The specific criteria and procedures used are described in Table S3 for aliphatics, and Table S4 for organic acids.

**Table S3.** Types of aliphatic peaks.

|  |  |  |  |
| --- | --- | --- | --- |
| *aliph.type* | two separated peaks | two unseparated peaks | one peak |
| *Image* |  |  |  |
| *Typical samples* | most peat and leaves | fresh *Sphagnum* | wood and paper |
| *Specific criteria* | The maximum absorbance of the aliphatic region is close to 2920, and there is a local minimum between 2850 and 2920. | The maximum absorbance of the aliphatic region is close to 2920. There is no local minimum between 2850 and 2920, but most of this region is below a straight line connecting these peaks (i.e., concave up). | The maximum absorbance of the aliphatic region is not close to 2920; or, it’s close to 2920, but most of the region between 2850–2920 is above a straight line connecting these points (i.e. concave down). |
| *Peak-finding procedure* | aliph29: Maximum of baseline-corrected absorbance between 2900 and w2.  trough28: Minimum of baseline-corrected absorbance between 2854 and aliph29.  aliph28: Maximum of baseline-corrected absorbance between w1 and trough28. | aliph29: Maximum of baseline-corrected absorbance between 2900 and w2.  trough28: Minimum of baseline-corrected absorbance between 2854 and aliph29.  aliph28: Use default value of 2850. | Since it is one peak, it is treated as such; i.e., aliph28, aliph29, and trough28 are identically assigned as maximum of the baseline-corrected absorbance between w1 and w2. |

An “aliphatic separation index” can be calculated as either aliph28/trough28 or aliph29/trough28, where higher values indicate greater peak separation. This index would equal 1 for one peak, and would be >1 for two peaks.

For the region containing the arom16, trough16, and acids, the arom16 peak is first found based on the maximum baseline-corrected absorbance between w1 and 1660. An initial guess for trough16 (tr.ini) is then found within an expected region (which varies depending on the locations of arom16 and w2) using procedures similar to finding baseline endpoints. The type of acids peak is then determined before finding its location and the location of trough16.

**Table S4.** Types of organic acid peaks.

|  |  |  |  |
| --- | --- | --- | --- |
| *acids.type* | peak | shoulder | no peak |
| *Image* |  |  |  |
| *Typical samples* | plants and acidic peat (e.g., bogs) | slightly acidic peat (e.g., northern fens) | slightly acidic to neutral peat (e.g., Everglades; very deep samples) |
| *Specific criteria* | tr.ini is a local minimum. | tr.ini is not a local minimum, but a “shoulder” acids peak can still be found based on where the second derivative above tr.ini is negative (concave down). | tr.ini is not a local minimum, and the second derivative above tr.ini is always positive (concave up), such that no acids peak can be found. |
| *Peak-finding procedure* | acids: Maximum of baseline-corrected absorbance above tr.ini or 1685 (whichever is greater).  trough16: Minimum of baseline-corrected absorbance between arom16 and acids. | acids: Minimum of second derivative above tr.ini or 1685 (whichever is greater).  trough16: Use tr.ini. | acids: Use default value of 1725.  trough16: Use default value of 1685. |

In addition to these designations, the label “negative” is applied to acids.type if the baseline-corrected absorbance of trough16 or acids (or both) is negative, as often occurs when the acids peak is extremely small relative to arom16 (causing the spectrum to dip below the baseline).

An “organic acid index” can be calculated as the ratio of acids/trough16, with higher values indicating higher organic acid content. If the peak type is labeled as “negative,” this index should not be calculated for the baseline-corrected peaks (which would then give unpredictable results), but it can still be calculated for peaks without the baseline corrections.