A computational protocol and software implementation (as a MATLAB application) for biomarker identification in infrared spectroscopy datasets

Julio Trevisan
Department of Communication Systems, Lancaster University, Lancaster LA1 4WA, UK; Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

Plamen P. Angelov
Department of Communication Systems, Lancaster University, Lancaster LA1 4WA, UK; Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

Francis L. Martin
Centre for Biophotonics, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK

DOI:
10.1038/nprot.2010.97

SUBJECT AREAS
Biochemistry *Computational biology and bioinformatics*

KEYWORDS
infrared, spectroscopy, biomarkers, feature selection, wavenumbers, exploratory analysis, pathways, mechanisms, computational framework, pattern classification, statistical significance, software, graphical user interface, visual representation, peak detection, differences between means, classification rates, importance histograms, bmtool, toolbox, toolkit, opensource, free software

Introduction

One attractive possibility of infrared (IR) spectroscopy is that it may be applied to investigate class (*i.e.*, treatment, tissue type, etc)-specific alterations in the absorption signature. Such alterations can act as biomarkers of mechanism associated with pathways or effects. One may be interested in investigating such alterations from different standpoints including: (1) intensity; (2) statistical significance; and, (3) composite (multi-spectral-region) alterations. These three view-concepts were implemented computationally and named BM1, BM2 and BM3. They can be easily applied to datasets of classed IR spectra through a user-friendly MATLAB interface.

BM1. Most intuitive of the methods. The mean spectrum from a given class is subtracted from the mean spectrum from a reference class (*e.g.*, “vehicle control”) thus obtaining a “difference-between-means curve”.

BM2. Each variable (wavenumber) is taken at a time as input to a univariate linear classifier thus obtaining a per-wavenumber “classification rate curve”[1]. Cross-validation is used to determine classification rates. This method is close to the t-test criterion[2], but more precise.

BM3. When multiple variables are assessed together, the joint-best variables for classification may differ substantially from the rank of the individual best variables. This method generates a histogram that represents how many times each wavenumber appeared within the *TopVars* (method parameter: number of “best variables”) “best variable set” achieved through feature selection, which is repeated many times according to *NoBootstraps* (method parameter: number of validation bootstraps).

The aim of this protocol is identify and visualize class-related biomarkers in IR spectral datasets by means of a simple sequence of steps to be executed under a user-friendly interface (**Figure 1**). Two visual representations are provided where all BM results are presented concurrently allowing for comparison of results generated by each method.

Equipment

Running this protocol will require:

MATLAB r2008a (www.mathworks.com) (version 7.6) or above. The toolkit is however probably compatible with older versions of MATLAB 7 (but not 6);
Our biospectroscopy toolkit, available for download at http://lec.lancs.ac.uk/biospectool;
A classed spectroscopy dataset (a sample data file is provided together with the toolkit).

Preparing data files

BMTool can read text files in CSV format (http://tools.ietf.org/html/rfc4180). A sample file is provided ("txt/vc_mnng.txt"), which can also be opened by Excel (www.microsoft.com), OpenOffice (www.openoffice.org) or a similar program. BMTool expects CSV files to be internally organized in the following way:

The whole file contains a matrix where each row represents one IR spectrum;
All columns, except for the last one, are real numbers representing absorption intensity values;
The last column is a positive integer specifying the class the spectrum belongs to. Class numbering can start with "0" or "1".

IR spectra represented in the file need to be pre-processed. Commonly-employed pre-processing sequences are: baseline correction followed by normalization to the Amide I peak[3,4] (or Amide II peak [5]) or second differentiation followed by vector normalization[6].

Alternatively, data can be imported from a spectroscopy database, a resource that is intended to be made publicly available in the future.

Procedure

1. Toolkit installation: 1.1. Download and extract the "bmtool.zip" file to a folder of choice. This will create five folders: "doc", "lib", "mat", "txt" and "work";

1.2. Start MATLAB;

1.3. Add the "lib" folder to your MATLAB path (File->Set path...->...).

2. Starting BMTool: 2.1. Start MATLAB and change the current directory to the "work" folder created upon installation;

2.2. At the MATLAB command prompt, type "bmtool". This will open the BMTool graphical user interface (GUI) (**Figure 2**).

3. Using bmtool - the GUI contains a set of panels that are numbered to facilitate the following operation steps. These steps go from loading a dataset to visualizing analysis results: 3.1. Load dataset:

3.1.1. In the "((1)) Select input dataset" panel, click on the "Load..." button;

3.1.2. Locate the desired file (for the sample file, by clicking "List files below" straight away this should show the sample data file “vc_mnng.txt” in the list box).

3.2. Mount datasets for analysis:

3.2.1. In the “((2)) Mount datasets for analysis” panel, make sure the “Agent vs. vehicle control (VC)” radio button is selected (it should be noted that Agent and VC are arbitrary labels);

3.2.2. Inform the index of the class that corresponds to the reference class (VC) in your dataset (1 for 1st, 2 for 2nd etc);

3.2.3. Click on “Mount datasets!” This will generate 2-class datasets where the VC data is combined with each of the remaining classes at a time. Consequently, the number of mounted datasets will be the number of classes existing in your dataset minus one.

3.3. Select datasets to be analysed: In the “((3)) Select mounted datasets” panel, select the desired datasets, or press the “Select all” button.

3.4. Generate results:

3.4.1. In the “((4)) Run session” panel, tick the “BM1 …”, “BM2 …” and “BM3 …” checkboxes.

Tip: recommended values:

for BM2: “Number of variables”: 1; “k for k-fold cross-validation”: 10;
for BM3: “Top number of variables”: 3; “Number of bootstraps”: 500.

3.4.2. Press the “Run!” button.

Caution: BM3 is much more time-consuming than the other methods and can be left out in a preliminary analysis that shows only results from BM1 and BM2. You can also run BM3 alone with a small number of bootstraps (*e.g.*, 20). This will give you an idea of the time taken to run the final result.

3.5. Visualize results:

3.5.1. In the “((5)) Manipulate results” panel, select which results you want to visualize or press the “Select all” button;

3.5.2. Select the type of visualization that you want to generate. Please note the availability of certain visualization types is conditioned to which results were generated:

To visualize the curves generated by each method, select “Mixed: BM curves” from the visualization options and click on “Visualize!” (see example in **Figure 3a**);
To visualize only the biomarker locations (without the curves), select “Mixed: Biomarker-localization

plots” from the visualization options and click on “Visualize!” (see example in **Figure 3b**);
To visualize the biomarkers in form of a report in MATLAB command window, select “Biomarker-localization report” and click on “Visualize!”.

Timing

Time is dependent on the computer setup, number of spectra, number of variables (*i.e.*, wavenumbers) in the dataset, and choice of method parameters. Times reported below result from the following settings: Intel Core i5-750 processor and a dataset conta

Troubleshooting

If BMTool does not run: verify that the “lib” folder was added to your MATLAB path.

If you cannot load the sample dataset: verify that your current working directory is the “work” folder.

Anticipated Results

Results from the sample data file are presented. The sample data file contains two treatment regimens (VC and *N*-nitroso-*N*-methylnitroguanidine (MNNG)) in Syrian hamster embryo (SHE) cells. The application of the BMs to this dataset allows one to investigate the effects of MNNG compared to corresponding control in SHE cells.

Figure 3a shows BM1, BM2 and BM3 curves with their respective five most important peaks marked. Peaks that are present in two different BM curves within a distance of 25cm^{-1} are connected by a dashed line that signifies confirmation of the importance of the respective IR region.

Figure 3b shows a compact version of the previous in a plot named *biomarker-localization* (BL) plot, where only the markers from Figure 3a are retained and symbol size is proportional to peak intensity. In addition, a BL plot containing seven comparisons of different treatment conditions vs. VC in a SHE study is shown in **Figure 3c**. The BL plot allows for quick visualization of biomarker weighting and comparison between classes in a class-rich dataset.

References

1. Martin, F.L. *et al*. Identifying variables responsible for clustering in discriminant analysis of data from infrared microspectroscopy of a biological sample. *J. Comput. Biol.*, **14** (9), 1176–84 (2007).

2. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine*

Learning Research, **3**, 1157–1182 (2003).

3. Walsh, M.J. *et al*. Fourier transform infrared microspectroscopy identifies symmetric PO2- modifications as a marker of the putative stem cell region of human intestinal crypts. *Stem Cells*, **26** (1), 108-18 (2008).

4. Kelly, J.G. *et al*. Derivation of a subtype-specific biochemical signature of endometrial carcinoma using synchrotron-based Fourier-transform infrared microspectroscopy. *Cancer Lett.*, **274** (2), 208-17 (2009).

5. German, M.J. *et al*. Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. *Biophys. J.*, **90** (10), 3783-95 (2006).

6. Matthäus, C. et. al. Infrared and Raman microscopy in cell biology. *Methods cell Biol.*, **89**, 275-308 (2009).

Acknowledgements

This work was funded by Unilever as “part of Unilever's ongoing effort to develop novel ways of delivering consumer safety”.

Figures

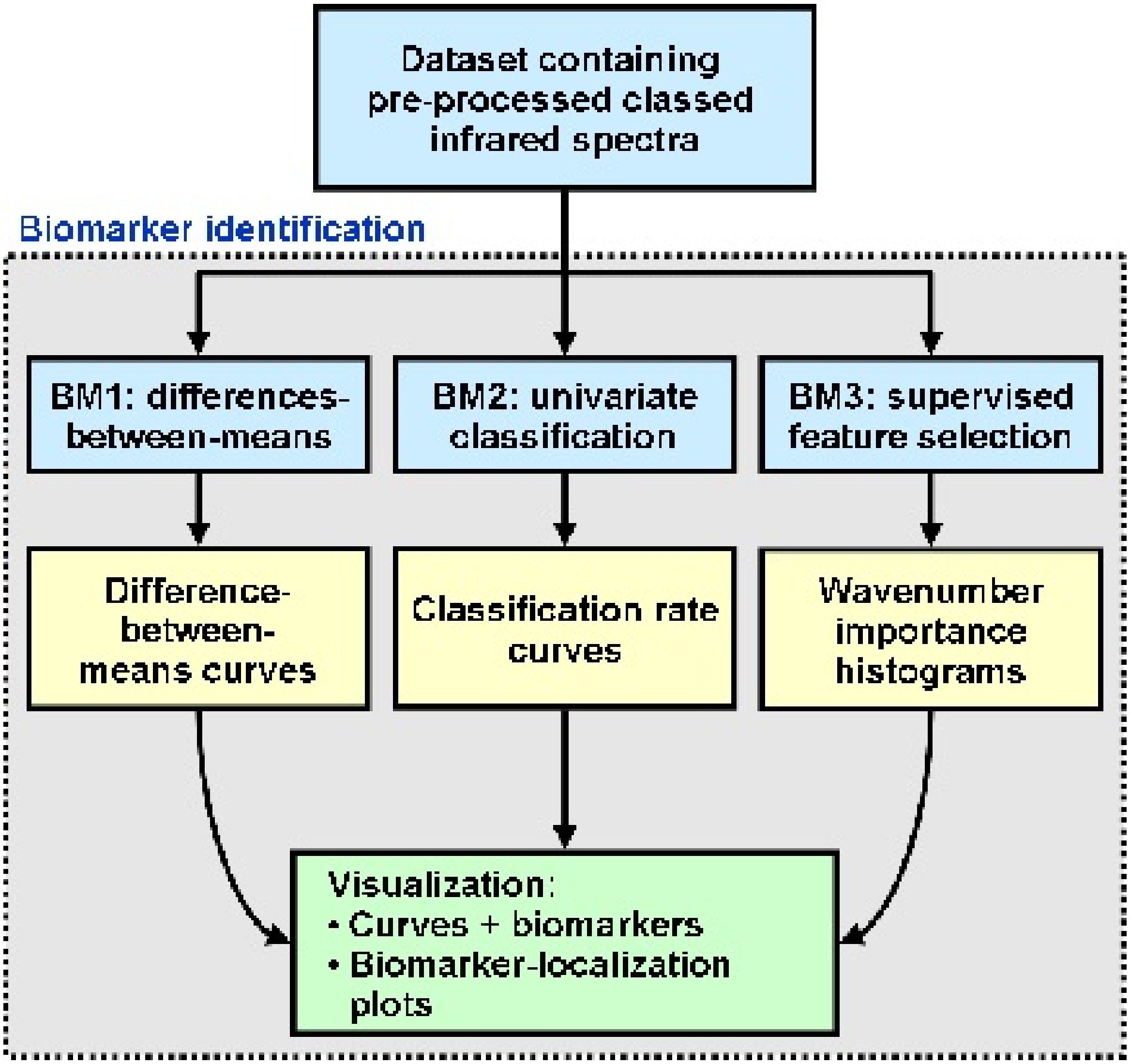
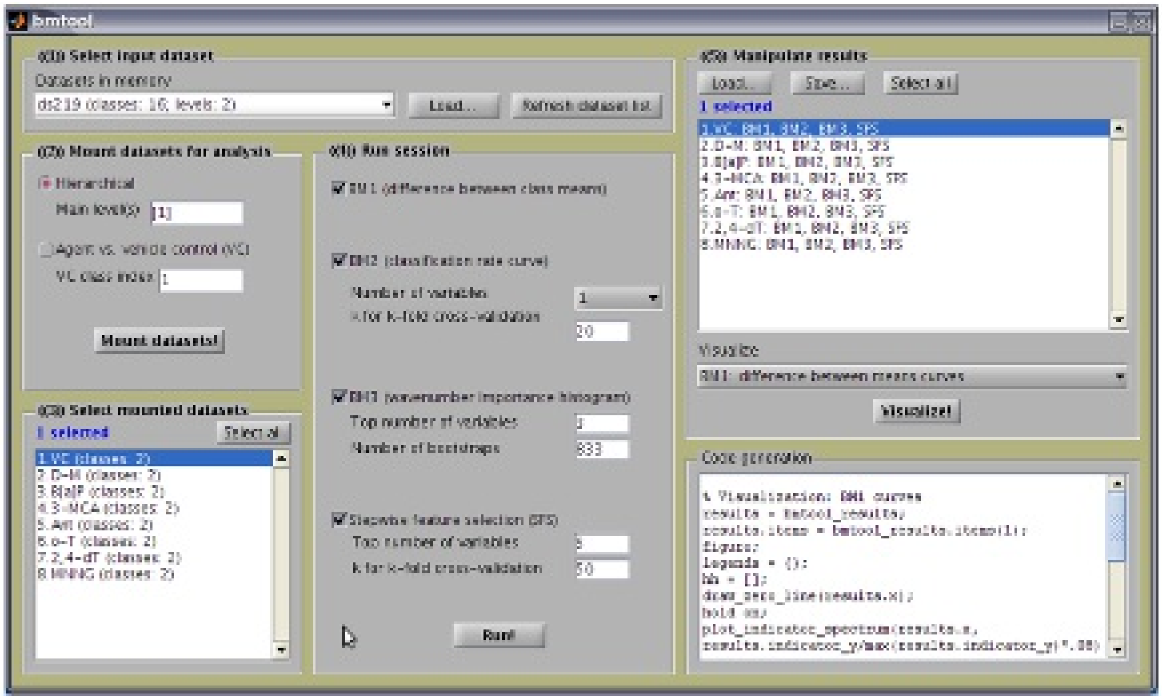

Dataset containing pre-processed classed infrared spectra
Biomarker identification
BM1: differences-between-means
BM2: univariate classification
BM3: supervised feature selection
Difference-between-means curves
Classification rate curves
Wavenumber importance histograms
Visualization:
• Curves + biomarkers
• Biomarker-localization plots

Figure 1

Schematic of the biomarker identification protocol implemented in the toolkit. Infrared spectra, pre-processed and classed (_i.e._, each spectrum is assigned a class, _e.g._, “vehicle control”, “treatment 1”, “treatment 2” etc), are inputted into three different biomarker identification methods (BM1, BM2 and BM3). Each method individually generates a result curve (see text). Results are combined by means of visualization strategies.



Datasets in memory: ds219 (classes: 16; levels: 2)

● Hierarchical
Main level(s): [1]
○ Agent vs. vehicle control (VC)
VC class index: 1

1 selected

1. VC (classes: 2)
2. D-M (classes: 2)
3. B[a]P (classes: 2)
4. 3-MCA (classes: 2)
5. Ant (classes: 2)
6. o-T (classes: 2)
7. 2,4-dT (classes: 2)
8. MNNG (classes: 2)

☑ BM1 (difference between class means)

☑ BM2 (classification rate curve)
Number of variables: 1
k for k-fold cross-validation: 20

☑ BM3 (wavenumber importance histogram)
Top number of variables: 3
Number of bootstraps: 833

☑ Stepwise feature selection (SFS)
Top number of variables: 5
k for k-fold cross-validation: 50

1 selected

1.VC: BM1, BM2, BM3, SFS
2.D-M: BM1, BM2, BM3, SFS
3.B[a]P: BM1, BM2, BM3, SFS
4.3-MCA: BM1, BM2, BM3, SFS
5.Ant: BM1, BM2, BM3, SFS
6.o-T: BM1, BM2, BM3, SFS
7.2,4-dT: BM1, BM2, BM3, SFS
8.MNNG: BM1, BM2, BM3, SFS

Visualize: BM1: difference between means curves

```
% Visualization: BM1 curves
results = bmtool_results;
results.items = bmtool_results.items(1);
figure;
legends = {};
hh = [];
draw_zero_line(results.x);
hold on;
plot_indicator_spectrum(results.x,
results.indicator_y/max(results.indicator_y)*.08)
```

Figure 2

Screenshot of the main window of BMTool. This graphical user interface contains numbered panels organized in the most probable operating sequence.

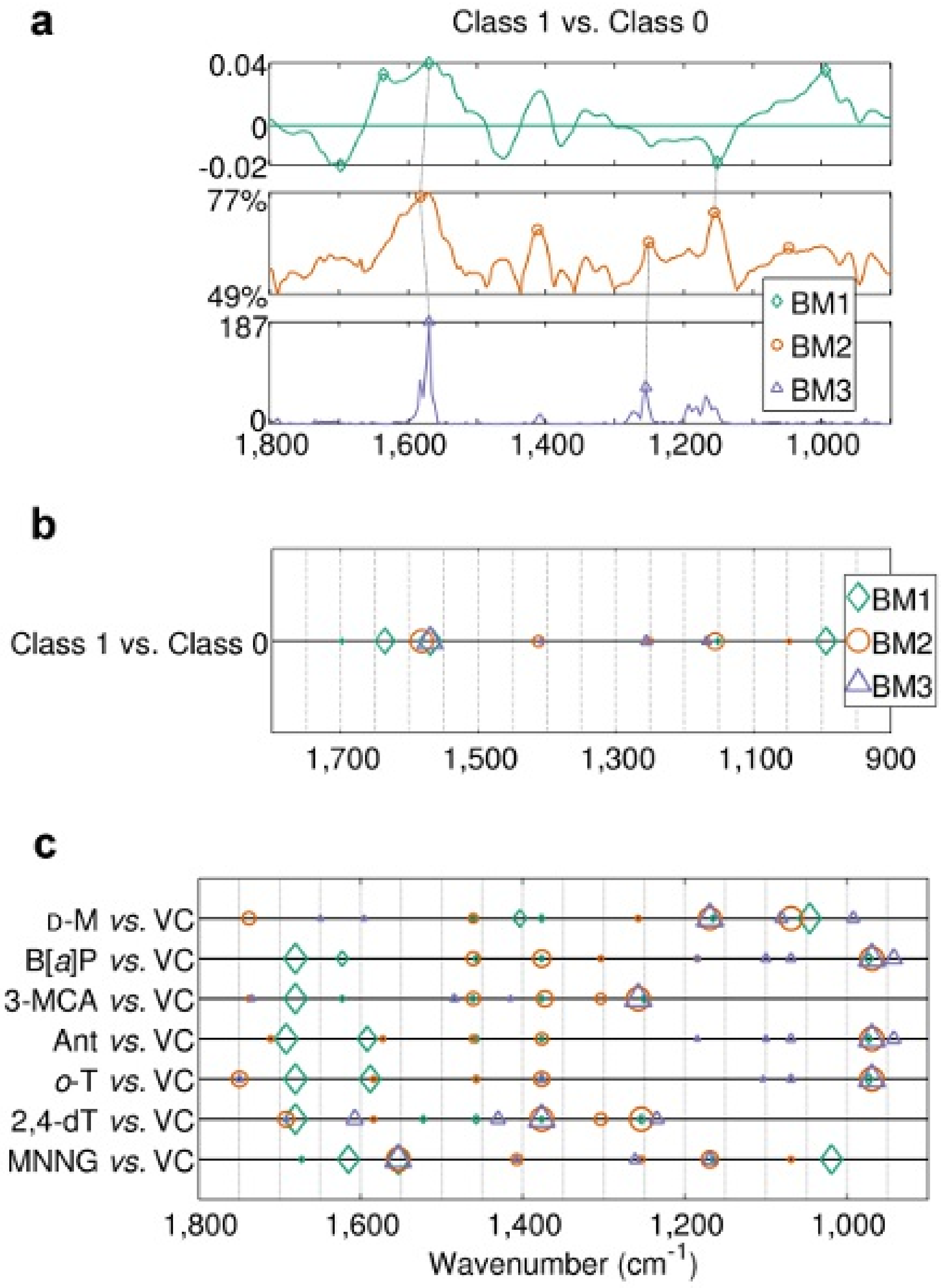


Figure 3

Anticipated results. _(a)_ BM1, BM2 and BM3 curves with their respective five most important peaks marked. Peaks that are present in two different BM curves within a distance of 25cm^-1^ are connected by a dashed line that signifies confirmation of the importance of the respective infrared region. _(b)_ Compact version of the previous in a plot named _biomarker-localization_ (BL) plot, where only the markers from Figure 3a are retained. _(c)_ Another BL plot containing all seven comparisons between treatment condition and VC of an eight-treatment-regimen Syrian hamster embryo (SHE) study; allows for quick visualization and comparison between classes in a class-rich dataset.