BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 2 2010, pages 280–282
doi:10.1093/bioinformatics/btp639

Systems biology

VIBE 2.0: Visual Integration for Bayesian Evaluation

Nathaniel Beagley[1], Kelly G. Stratton[2] and Bobbie-Jo M. Webb-Robertson*

[1]Computational Mathematics and [2]Computational Biology and Bioinformatics,
Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received on April 28, 2009; revised on October 5, 2009; accepted on November 10, 2009

Advance Access publication November 17, 2009

Associate Editor: Trey Ideker

ABSTRACT

Summary: Data fusion methods are powerful tools for evaluating experiments designed to discover measurable features of directly unobservable systems. We describe an interactive software platform, Visual Integration for Bayesian Evaluation, that ingests or creates Bayesian posterior probability matrices, performs data fusion and allows the user to interactively evaluate the classification power of fusing various combinations of data sources, such as transcriptomic, proteomics, metabolomics, biochemistry and function.

Availability: http://omics.pnl.gov/software/VIBE.php

Contact: bj@pnl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The goal of data fusion is to integrate different types of data about a system to create models that are more complete and accurate than those derived from any individual data source, i.e. the whole is greater than the sum of its parts. The improvement of statistical classification, and direct applications such as prediction of protein function, is often the end goal of data fusion; however, heterogeneity of the data (varying dynamic range and specificity) presents a major challenge. Methods that transform the data into a common form, such as kernel matrices or Bayesian posterior probabilities, are often used (Hwang *et al.*, 2005; Jarman *et al.*, 2008; Lanckriet *et al.*, 2004; Troyanskaya *et al.*, 2003; Webb-Robertson *et al.*, 2009), since after applying such methods, fusion is simply a matter of merging matrices in a statistical manner.

Fusion methods look to take advantage of orthogonal information captured by multiple analytical platforms that, when taken together, increase the classification power over that of a single measurement platform, thus allowing more accurate and complete predictions of the phenotype of interest. However, in many cases improvement is only achieved with a subset of the available data sources (Lu *et al.*, 2005), and therefore it is important to provide an intuitive interpretation of these results in an interactive form that can be used to evaluate the impact of each individual data stream in the context of the overall integrated analysis. Visual Integration for Bayesian Evaluation (VIBE) 2.0 offers a simple and flexible approach to fuse complementary datasets and dynamically evaluate the contribution of each dataset.

VIBE 2.0 is a stand-alone software tool that allows a user to explore the effects of including or excluding specific data sources in a Bayesian fusion analysis. VIBE works by integrating probability models from multiple data streams. The software can either ingest precomputed probability models or create them from the raw data. The statistical methods used to derive the probability models and the data that is included in the fusion can be modified on the fly to analyze the system dynamically.

2 ANALYSIS CAPABILITIES

VIBE 2.0 takes as input either raw datasets or precomputed probability models for each data type. The probability model is a matrix where each value (i,j) is the probability of observing a specific known experimental group (j) given a sample (i) associated with one or more datasets. To create these probability models VIBE uses statistical learning algorithms, including naïve Bayes classification (Mitchell, 1997), degree of association (Jarman *et al.*, 2000), k-nearest neighbors (Ativa, 2005) and multinomial logistic regression (McCullagh and Nelder, 1990). These statistical learning algorithms compute the probability of observing the specific data associated with a sample given a particular experimental group. Bayesian statistics are used to generate the posterior probabilities that are represented in the probability matrices used by VIBE (Webb-Robertson *et al.*, 2009). For each data source, VIBE 2.0 then calculates the classification accuracy (the fraction of samples assigned to the correct experimental group) providing the user a baseline that shows quantitatively the effectiveness of each individual analysis platform. A class assignment table is also graphically displayed, depicting the experimental groups into which the true samples from an experimental group are classified. The visualization allows the user to gain insight into the efficacy of the individual platforms, for example, showing that a particular data type is unable to distinguish between two of the experimental groups. The user then selects a subset (or the full set) of the data sources to be included in the integrated analysis and VIBE 2.0 performs a Bayesian fusion and gives the classification accuracy based on the integrated probability model (Webb-Robertson *et al.*, 2009). As the fusion calculation is almost instantaneous, the user can experiment with multiple combinations of the input data sets to evaluate the impact of including each data set in the fused analysis.

*To whom correspondence should be addressed.

3 IMPLEMENTATION

VIBE 2.0 is built in MATLAB® 2009b from The Mathworks, Inc.® and is packaged, using Version 4.11 MATLAB® Compiler, as a stand-alone executable for the Windows platform. The application consists of three graphical user interface screens. The first two screens are associated with the user 'input' where the data sources are specified, as well as the statistical methods that will be used to analyze each dataset. The third or 'analysis' screen then provides visualization of the data integration analysis.

3.1 Input screen

On the input screen (Fig. 1A), the user specifies the source data files containing the class matrix (defining the true experimental group of each sample in the experiment) and the raw data or probability matrices for each individual data source to be used in the integration. The data handling screen (Fig. 1B) is used to select the type of data for upload and the statistical method to be used to create the probability model. VIBE 2.0 does not perform any data quality checks beyond assuring dataset sample sizes match and the data have appropriate values for the statistical method to be employed. The assumption is that the data are of adequate quality and has been properly normalized prior to analysis. VIBE 2.0 does offer auto-scaling of the data, which will normalize all variables to have a common mean of zero and unity variance. The uploaded files can be MATLAB® (.mat), Microsoft Excel (.xls or .xlsx) or flat text (.txt) files. There are also fields in the input screen where the user may also enter a name, an abbreviation and a brief description for each dataset, as well as names for each experimental group if they are not specified in the class file. Once all information is entered, the user presses the 'Continue' button to launch the 'analysis' screen.

3.2 Analysis screen

The analysis screen (Fig. 1C) has two visualization sections, one displaying the analysis results of each individual data source and a second displaying the results of the data fusion. Although the example shown has three data sources, up to six are viewable simultaneously. Upon launch, the classification accuracy and the class assignment table for each individual data source are calculated and displayed. The class assignment table is displayed as a plot with true class along the left axis and predicted class along the top axis, where the color at each location represents the fraction of samples classified into the associated class. Thus, a diagonal line of red boxes running top left to bottom right represents perfect classification.

The user selects the subset of datasets to use in the integrated analysis via the 'Use in Integration' buttons adjacent to each data source (default is all selected). The 'Integrate' button calculates the integrated probability model and displays the classification accuracy and the class assignment table for the fused analysis. Multiple combinations can be explored interactively as the calculation is nearly instantaneous.

Additional features are available to facilitate the use of the integrated results in further analysis. Optional annotations can be added and the 'Save Screen' button saves a jpeg image of the current state of the analysis screen. The 'Output File' button exports a (.xls) file containing results from the integrated analysis giving, for each sample in the experiment, the true class, the predicted class and probability of being assigned to the predicted class.

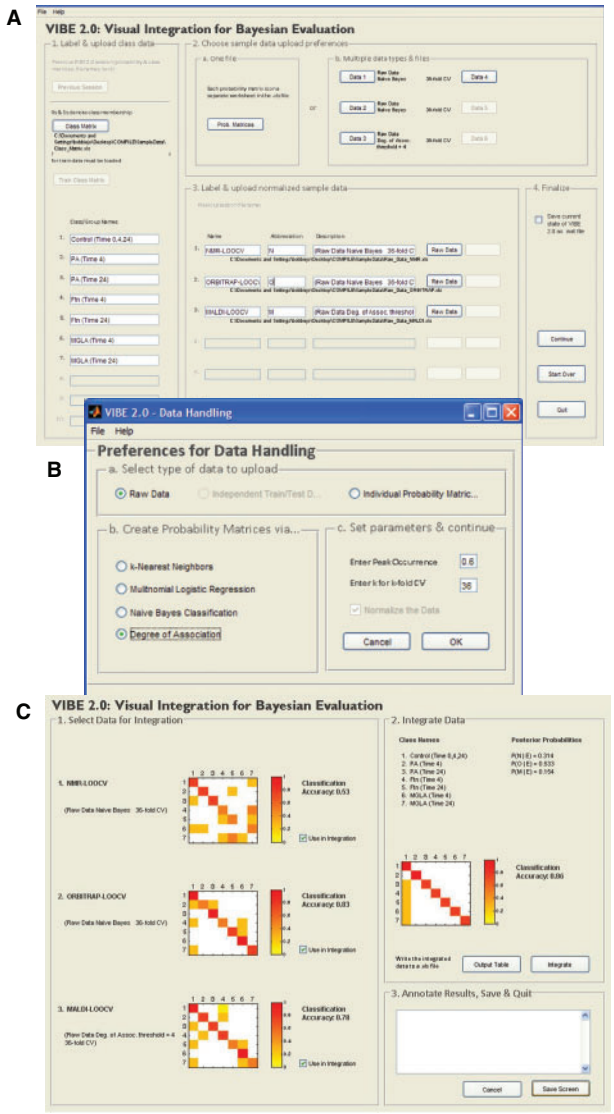

Fig. 1. (**A**) The input screen is shown as it appears with data loaded for fusion. (**B**) Shows the data handling screen for the third dataset (MALDI) for leave-one-out cross-validation. (**C**) Output analysis screen showing the results of integrating all three datasets.

4 CASE STUDY

A previously described experiment to detect early response in mice to *Francisella novicida* (FTN) is shown in Figure 1 (Webb-Robertson *et al.*, 2009). The experiment shows seven classes where the mice are exposed to one of three microbes at both 4 and 24 h; FTN, *Pseudomonas aeruginosa* (PA), or an avirulent strain of FTN that contains a mutation to the transcriptional regulator mglA (MGLA). Bronchial alveolar lavage fluid was collected from each animal and analyzed using three instrument platforms: nuclear magnetic resonance spectroscopy (NMR), matrix assisted laser desorption/ionization mass spectrometry (MALDI) and accurate mass and time mass spectrometry (Orbitrap)™. Features were extracted and a probability model was constructed for each instrument using either naïve Bayes classification (Mitchell, 1997)

or degree of association (Jarman *et al.*, 2000). The probability matrices as input to VIBE can either be the result of independent test data or the result of cross-validation, as is the case for this example. Details of this analysis can be found in the user manual available through the software.

VIBE 2.0 was used to explore the metabonomics and proteomics results using different combinations of the three instruments in an integrated analysis. As demonstrated in Figure 1, a higher level of classification accuracy is achieved by using all three datasets than can be achieved from any one individual dataset. This example also demonstrates that incorporating data from additional instruments does not always improve results. The probability models were developed using leave-one-out cross-validation, which is equivalent to the number of separations as samples in the data (Fig. 1B). The classification accuracy of using NMR and MALDI is 61% compared with 78% using MALDI alone (data not shown). Similarly, classification accuracy is 81% with MALDI and Orbitrap™ compared with 83% with Orbitrap™ alone (data not shown), suggesting that MALDI analysis does not complement the NMR and Orbitrap™ datasets as might have been expected. However, the integration of only NMR and Obitrap attains an accuracy of 86%, which is the same as integrating all three datasets (Fig. 1C).

ACKNOWLEDGEMENTS

PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RL01830.

Funding: U.S. Department of Energy through the Environmental Biomarkers Initiative at Pacific Northwest National Laboratory; National Institutes of Health (grants U54 016015 and U54 AI057141).

Conflict of Interest: none declared.

REFERENCES

Atiya,A.F. (2005). "Estimating the posterior probabilities using the k-nearest neighbor rule." *Neural Comput.*, **17**, 731–740.

Hwang,D. *et al.* (2005) A data integration methodology for systems biology. *Proc. Natl Acad. Sci. USA.* **102**, 17296–17301.

Jarman,K.H. *et al.* (2000) An algorithm for automated bacterial identification using matrix-assisted laser desorption/ionization mass spectrometry. *Anal. Chem.*, **72**, 1217–1223.

Jarman,K.H. *et al.* (2008) Bayesian-integrated microbial forensics. *Appl. Environ. Microbiol.*, **74**, 3573–3582.

Lanckriet,G.R. *et al.* (2004) A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626–2635.

Lu,L.J. *et al.* (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.

McCullagh,P. and Nelder,J.A. (1990). *Generalized Linear Models*. Chapman & Hall, New York.

Mitchell,T. (1997). *Machine Learning*. McGraw Hill Higher Education, Columbus.

Troyanskaya,O.G. *et al.* (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction. *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.

Webb-Robertson,B.-J. *et al.* (2009) A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections. *Pac. Symp. Biocomput.*, **14**, 451–463.