The scale of protein-based measurements is rapidly increasing with the advent of so-called ‘multiplexing’ technologies for antibody-based measurement offered by platforms such as Luminex and Meso-Scale Discovery. These technologies are readily amenable to intricate experimental layouts which simultaneously investigate multiple experimental conditions, treatments and time points simultaneously. Examples of such complex and large-scale protein-based datasets are starting to appear in the literature (Cosgrove et al., 2005; Mitsos et al., 2008). For cases where measurements are made in a combinational fashion across multiple dimensions, there is a need for a tool to efficiently manipulate and reorganize such data for visualization. DataPflex accepts data consisting of up to five arbitrary dimensions in addition to a measurement dimension. Data are imported from a simple .xls format and can be exported to MATLAB or .xls. Data dimensions can be reordered, subdivided, merged, normalized and visualized in the form of collections of line graphs, bar graphs, surface plots, heatmaps, ICG50’s and other custom plots. Open source implementation in MATLAB enables easy extension for custom plotting routines and integration with more sophisticated analysis tools.

Motivation: DataPflex is a MATLAB-based application that facilitates the manipulation and visualization of multidimensional datasets. The strength of DataPflex lies in the intuitive graphical user interface for the efficient incorporation, manipulation and visualization of high-dimensional data that can be generated by multiplexed protein measurement platforms including, but not limited to Luminex or Meso-Scale Discovery. Such data can generally be represented in the form of multidimensional datasets (e.g. time × inhibitor × inhibitor concentration × cell type × measurement). For cases where measurements are made in a combinational fashion across multiple dimensions, there is a need for a tool to efficiently manipulate and reorganize such data for visualization. DataPflex accepts data consisting of up to five arbitrary dimensions in addition to a measurement dimension. Data are imported from a simple .xls format and can be exported to MATLAB or .xls. Data dimensions can be reordered, subdivided, merged, normalized and visualized in the form of collections of line graphs, bar graphs, surface plots, heatmaps, ICG50’s and other custom plots. Open source implementation in MATLAB enables easy extension for custom plotting routines and integration with more sophisticated analysis tools.

Availability: DataPflex is distributed under the GPL license (http://www.gnu.org/licenses/) together with documentation, source code and sample data files at: http://code.google.com/p/datapflex.

Contact: DataPflexinfo@gmail.com; bhendriks@merinackpharma.com

Supplementary information: Supplementary data available at Bioinformatics online.

1 INTRODUCTION

The scale of protein-based measurements is rapidly increasing with the advent of so-called ‘multiplexing’ technologies for antibody-based measurement offered by platforms such as Luminex and Meso-Scale Discovery. These technologies are readily amenable to intricate experimental layouts which simultaneously investigate multiple experimental conditions, treatments and time points simultaneously. Examples of such complex and large-scale protein-based datasets are starting to appear in the literature (Cosgrove et al., 2008; Mitsos et al., 2009; Samaga et al., 2009). For cases where measurements are made in a combinational fashion across multiple dimensions, there is a need for a tool to efficiently manipulate and reorganize such data for visualization. DataPflex accepts data consisting of up to five arbitrary dimensions in addition to a measurement dimension. Data are imported from a simple .xls format and can be exported to MATLAB or .xls. Data dimensions can be reordered, subdivided, merged, normalized and visualized in the form of collections of line graphs, bar graphs, surface plots, heatmaps, ICG50’s and other custom plots. Open source implementation in MATLAB enables easy extension for custom plotting routines and integration with more sophisticated analysis tools.

Availability: DataPflex is distributed under the GPL license (http://www.gnu.org/licenses/) together with documentation, source code and sample data files at: http://code.google.com/p/datapflex.

Contact: DataPflexinfo@gmail.com; bhendriks@merinackpharma.com

Supplementary information: Supplementary data available at Bioinformatics online.

2 METHODS AND IMPLEMENTATION

DataPflex was written entirely in MATLAB R2007b (The Mathworks, Natick, MA, USA; the graphical user interface has known incompatibility with prior MATLAB releases). All experimental data must be categorized into six dimensions: (i) arbitrary dimensions and a measurement dimension. Each dimension can be either textual (inhibitor names, for example) or numeric (time points or concentrations, for example) in nature. For dataset in excess of six dimensions, it is recommended that such datasets be split into multiple input files.

Data is imported from an Excel (.xls) file, consisting of two blocks separated by an empty column. The first block is a description block with up to five dimension headings of the form ‘dimension name=’ or ‘dimension name: dimension units’. For cases where there are pairs of corresponding

Bart S. Hendriks*† and Christopher W. Espelin
Pfizer Research Technology Center, 620 Memorial Drive, Cambridge, MA 02139, USA

Associate Editor: Trey Ideker

Received on October 13, 2009; revised on November 14, 2009; accepted on November 27, 2009
Advance Access publication December 4, 2009

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Table 1. Example Excel (.xls) data format #1 for import into DataPflex (individual replicates)

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Stimali</th>
<th>Time/min</th>
<th>Inhibitor</th>
<th>Inhibitor Concentration=μM</th>
<th>Measurement #1=units</th>
<th>Measurement #2=units</th>
<th>…</th>
</tr>
</thead>
</table>

[Data Description Block]

[Experimental Data Block]

![Fig. 1. DataPflex screenshot.](https://academic.oup.com/bioinformatics/article-abstract/26/3/432/214374)

Dimensions (inhibitor and inhibitor concentration, for example), dimension headings should be [name] and [name] concentration = units, where [name] is the name of the dimension (inhibitor and inhibitor concentration = μM, for example). These column headings may be in any order and the description block may contain an arbitrary number of additional columns with empty headings. The second block is the measurement data, separated from the description block by an empty column. The data block can be in one of two formats: (Table 1) with column headings correspond to the measurements performed or (Supplementary Material) with column heading corresponding to the mean and standard deviations for each measurement. In all column headings, the unit description (=units) is optional. The first row of the file must contain the headings and subsequent rows contain corresponding experimental descriptions and data. For import format #1 (Table 1) replicates are entered as in separate rows and automatically averaged upon import with SDs calculated. A sample dataset from Espelin et al. (C.W.Espelin et al., manuscript in preparation) is provided in the Excel import format.

Following import, the six dimensions are shown within the DataPflex interface and can be efficiently and easily reordered with drop-down menus and subportions selected in each window (Fig. 1). Individual labels can be renamed and reordered within a given dimension. Multiple normalization options are offered for plotting, including raw data, normalization by the max value, log normalization, normalization relative to conditions/time courses within the dataset and several others. Additionally, the data corresponding to ‘No Treatment’ can be automatically added to each plot (bottom left of GUI, Fig. 1). Allowing, for example, the comparison of a treatment condition with background. Once the dimensions and data normalization have been selected, several plot types may be generated: line graphs, scatter plots, bar graphs, 3-D surface plots, mean/standard deviation, box plots, and heatmaps and arrow plots. Mean/Stdve plots display the mean and standard deviation of each selected measurement in the dataset in order to obtain a general view of which species in the dataset are showing significant variability.

Additional options exist for each plot type: (i) data can be clustered by row and/or column; and (ii) additional dimensions can be ‘unfolded’ (i.e. concatenated into either the rows or columns. For heatmaps, when the third dimension is a numeric dimension, such as time or concentration, the average value is plotted and the figure can be clicked on to visualize the underlying time series data. Line arrow plots contain the same information as heatmaps, but are presented in the form of an influence diagram where the thickness and color of each arrow indicates the strength of interaction. Arrow plots can also be clustered to improve presentation. A slider is provided to filter out weak interactions and entries not connected by any arrows are removed from the plots. Circle arrow plots are an analogous plot in which entries are plotted in a circle, enabling visualization of interactions within a network (stimulating with cytokines and measuring their release, for example). For each plot type, multiple color schemes are available. Sample plots from data presented in C.W.Espelin et al. (manuscript in preparation) are shown in the DataPflex Reference Guide (Supplementary Material). IC50 plots and clustering require the Statistics’ Toolbox (The Mathworks) and optimal leaf ordering for clustering relies on the Bioinformatics Toolbox (The Mathworks, Natick, MA).

Finally, dataset manipulation and import/export capability are also integrated: Users may (i) rename datasets, (ii) create a new dataset that is a subset of another dataset, (iii) filter or split a dataset into two based on user-defined criteria; and (iv) merge datasets. In addition to: .xls files, datasets can be imported and exported from either the MATLAB workspace or .mat files. Lastly, individual datasets (with both mean and standard deviation data) can be rewritten back to the Excel import .xls format and are ready for immediate reinput into DataPflex.

ACKNOWLEDGEMENTS

The authors would like to thank Arthur Goldsipe, Julio Saez-Rodriguez, Leo Alexopoulos, Natalia Boskharov, Deb Nathan, Marya Liimatta, Lucy Sun, Fei Hua, Nidhi Jhawar, Rocco Coli and Romesh Subramanian for their helpful discussions, application testing and feature ideas.

Funding: This work was funded by Pfizer, Inc.

Conflict of Interest: B.S.H. and C.W.E. hold stock in Pfizer, Inc.

REFERENCES


