VizRank: Finding Informative Data Projections in Functional Genomics by Machine Learning

(Application Note)

Gregor Leban[1], Ivan Bratko[1,2], Uros Petrovic[2] , Tomaz Curk[1] and Blaz Zupan[1,3,*]

[1] University of Ljubljana, Faculty of Computer and Information Science, [2] Jozef Stefan Institute, Ljubljana, Slovenia, [3] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA

ABSTRACT

Summary: VizRank is a tool that finds interesting two-dimensional projections of class-labeled data. When applied to multi-dimensional functional genomics data sets, VizRank can systematically find relevant biological patterns.

Availability: http://www.ailab.si/supp/bi-vizrank

Contact: blaz.zupan@fri.uni-lj.si

Supplementary information: http://www.ailab.si/supp/bi-vizrank

* To whom correspondence should be addressed

INTRODUCTION

In the study of gene function and gene interactions, functional genomics relies on various data analysis approaches. These include classification methods which assume that data for each gene consist of experimental measurements and a class label that associates the gene to some group of interest. These classes may represent gene functional categories, results of clustering, or any grouping of genes for which an expert believes that there is an inherent relationship. Various techniques for data visualization (McCarthy *et al.*, 2004) may complement or even provide an alternative to computational methods for inference of classification models (*e.g.*, support vector machines (Brown *et al.*, 2000)) to search for biologically interesting patterns. We show that even simple visualization techniques, such as a scatterplot, may be fitted for this task, provided that it visualizes the right subset of features included in the data. In functional genomics, finding such feature subsets is not trivial, since in a typical gene expression assay several tens or hundreds of measurements may be recorded for each gene at different experimental conditions, and manual search for interesting data projections is not practical.

Here we describe VizRank, a tool that automatically ranks and discovers interesting two-dimensional projections of class-labeled data. To see how VizRank can discover relevant biological patterns from functional genomics data, we considered an example on the budding yeast *Saccharomyces cerevisiae* data studied by Brown *et al.* (2000) where each gene is described by 79 different DNA microarray hybridization measurements. While this particular data set includes normalized log expression ratio measurements, VizRank can consider any type of continuous data for which the user is interested in finding meaningful visualizations. To show Brown *et al.*'s data in a two-dimension scatterplot, $79 \times 78/2 = 3081$ different projections

are possible. We used the data on three functional groups – respiration (30 genes), cytoplasmic ribosomes (121 genes), and proteasome (35 genes), and evaluated the scatterplots with VizRank. The scatterplot with the highest VizRank score (Figure 1.b) shows that measurements during sporulation and diauxic shift clearly separate the three functional groups. Gene expression during diauxic shift can characterize two out of the three functional groups – cytoplasmic ribosomes and respiration – which has already been reported (DeRisi *et al.*, 1997). A measurement during sporulation is required to clearly separate these two groups from the proteasome group. Only five out of 3081 projections (less than 0.2%) provide group discrimination as clear as the described scatterplot. For comparison, Figure 1.c shows a scatterplot with an average VizRank score. Interestingly, while reporting that separation of functional groups is possible by support vector machine classifier, Brown *et al.* (2000) – probably due to the difficulty of the interpretation of the classifier – did not report on particular rules that characterize the functional groups. As pointed out by the VizRank scatterplot, such rules do exist and could be easily visualized and interpreted.

AUTOMATIC RANKING OF PROJECTIONS

Given a data set where instances are described with N features, a geometric two-dimensional data projection P is a mapping $<\mathbf{x}, \mathbf{y}> = P(<\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N>)$, where $<\mathbf{x}, \mathbf{y}>$ are vectors with coordinates of projected data points and $\mathbf{v}_i$ are vectors with original features. The class label of data instances is mapped to the color or shape of the visualized point, and we are interested in projections with good class separation (*cf.* Figure 1.b). In an interesting projection, an instance would be surrounded by many instances of the same class. Following this observation and to score a specific projection, VizRank employs the k-nearest neighbor (k-NN) classifier – a

machine learning algorithm that when classifying an example finds its *k* nearest neighbors and classifies it to the prevailing class (*e.g.*, Mitchell, 1997). The score of the projection is estimated as the classification accuracy of the *k*-NN classifier evaluated on all data instances in the projected space. This scoring function is a good estimate of projection usefulness since projections with well-separated classes would be associated with high classification accuracy whereas projections with overlapping classes would score lower. Other machine learning methods could also be used, but we found the *k*-NN appropriate since it is insensitive to the shape and orientation of the class clusters.

VizRank can be applied to any visualization method that maps data to points in a two-dimensional space. Besides with scatterplot, we have also implemented it with radviz (Hoffman *et al.*, 1997; see supplement for further information) that can visualize an arbitrary number of features and use a nonlinear mapping of high-dimensional space to two dimensions.

By evaluating plots that use the original, untransformed set of features from experimental measurements and providing a ranked list of projections, VizRank compares favorably to other popular projection search methods such as principal component and discriminant analysis. For more detailed comparison and a heuristic approach that helps VizRank to find top-rated projections by evaluating only a small subset of possible projections see supplemental information.

IMPLEMENTATION

VizRank is implemented within an open-source data mining suite called Orange (Demsar and Zupan, 2004). Figure 1.a shows a snapshot of a part of VizRank's graphical interface. Detailed description of the interface and a web based VizRank demo is available in the supplement.

ACKNOWLEDGEMENTS

We thank Gad Shaulsky for discussions and comments on the paper. This work was supported in part by a grant from the Slovene Ministry of Education, Science and Sports and by a grant from the National Institute of Child Health and Human Development, P01 HD39691.

REFERENCES

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C., Furey,T.S., Ares,M., Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences*, **1**, 262–267.

Demsar,J., Zupan,B. (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, A White Paper. AI Lab, Faculty of Computer and Information Science, Ljubljana.

DeRisi J, Iyer V, Brown P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–6.

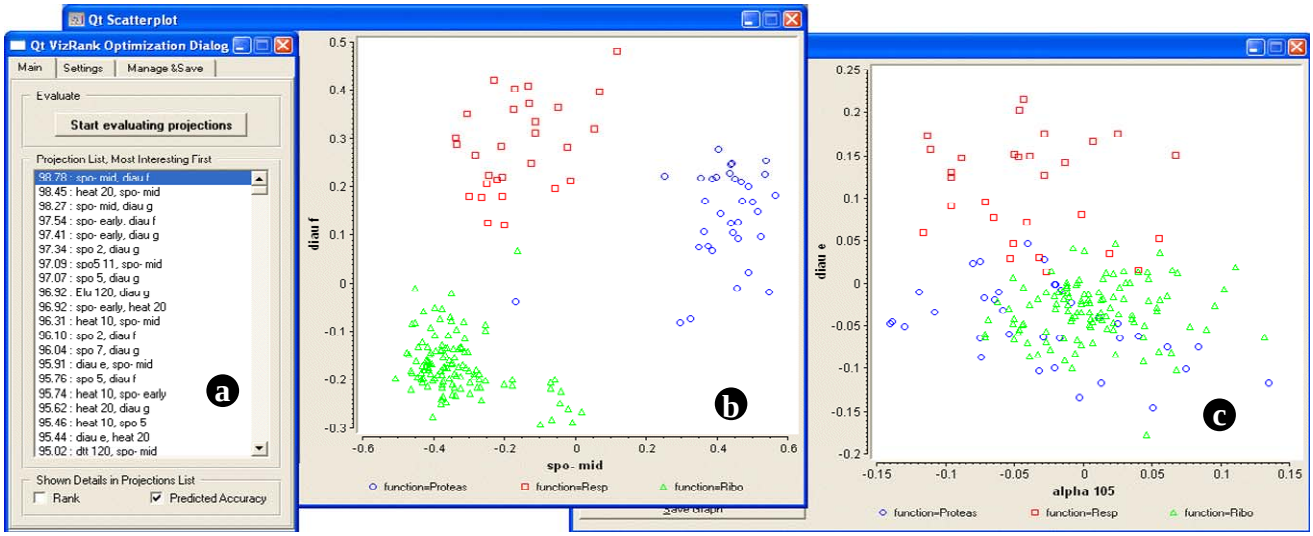
McCarthy,J.F., Marx,K.A., Hoffman,P.E., Gee,A.G., O'Neil P., *et al.* (2004) Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management, *Annals of the New York Academy of Sciences*, **1020**, 239–262.

Hoffman,P.E., Grinstein,G., Marx,K., Grosse,I., Stanley,E. (1997) DNA Visual and Analytic Data Mining, *IEEE Visualization 1997*, **1**, 437–441.

Figure legend

Figure 1: Snapshot of the VizRank dialog (a) and two scatterplots (b, c) from *S. cerevisiae* data studied in Brown *et al.* (2000). Using the default parameters, VizRank assigned a score of 98.78 (in the scale from 0 to 100) to the left, and a score of 72.50 to the right scatterplot.

Figure 1:


Qt Scatterplot
Qt VizRank Optimization Dialog
Main
Settings
Manage &Save
Evaluate
Start evaluating projections
Projection List, Most Interesting First
98.78 : spo- mid, diau f
98.45 : heat 20, spo- mid
98.27 : spo- mid, diau g
97.54 : spo- early, diau f
97.41 : spo- early, diau g
97.34 : spo 2, diau g
97.09 : spo5 11, spo- mid
97.07 : spo 5, diau g
96.92 : Elu 120, diau g
96.92 : spo- early, heat 20
96.31 : heat 10, spo- mid
96.10 : spo 2, diau f
96.04 : spo 7, diau g
95.91 : diau e, spo- mid
95.76 : spo 5, diau f
95.74 : heat 10, spo- early
95.62 : heat 20, diau g
95.46 : heat 10, spo 5
95.44 : diau e, heat 20
95.02 : dtt 120, spo- mid
a
Shown Details in Projections List
Rank
Predicted Accuracy
diau f
spo- mid
function=Proteas
function=Resp
function=Ribo
b
diau e
alpha 105
function=Proteas
function=Resp
function=Ribo
c