Bioinformatics, 32(3), 2016, 465–468
doi: 10.1093/bioinformatics/btv591
Advance Access Publication Date: 14 October 2015
Applications Note

Gene Expression

BiTrinA—multiscale binarization and trinarization with quality analysis

Christoph Müssel[1,†], Florian Schmid[1,†], Tamara J. Blätte[1,2], Martin Hopfensitz[1], Ludwig Lausser[3,‡] and Hans A. Kestler[1,3,*,‡]

[1]Medical Systems Biology, Ulm University, 89069 Ulm, Germany, [2]Section of Oncology, Internal Medicine III, Ulm University, 89069 Ulm, Germany and [3]Leibniz Institute on Aging—Fritz Lipmann Institute, 07745 Jena, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡These authors are joint senior authors.

Associate Editor: Janet Kelso

Received on March 10, 2015; revised on September 30, 2015; accepted on October 8, 2015

Abstract

Motivation: When processing gene expression profiles or other biological data, it is often required to assign measurements to distinct categories (e.g. 'high' and 'low' and possibly 'intermediate'). Subsequent analyses strongly depend on the results of this quantization. Poor quantization will have potentially misleading effects on further investigations. We propose the `BiTrinA` package that integrates different multiscale algorithms for binarization and for trinarization of one-dimensional data with methods for quality assessment and visualization of the results. By identifying measurements that show large variations over different time points or conditions, this quality assessment can determine candidates that are related to the specific experimental setting.
Availability and implementation: BiTrinA is freely available on CRAN.
Contact: hans.kestler@leibniz-fli.de or hans.kestler@uni-ulm.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the context of analyzing biological data, such as gene expression profiles, it is often necessary to identify those genes, transcripts or proteins that display prominent variations over time or across different experimental conditions. This can be required to preselect more interesting candidates for subsequent analyses or to reveal data applicable for further approaches, such as Boolean network modeling. This can be achieved by categorizing the data to a small number of distinct groups and assigning validities to this categorization.

For simplicity, we will focus in the following on binarization schemes and mention trinarization as their extension in the next section. Binarization is the process of partitioning data into two groups and assigning a single value to each. It can be seen as a special case of discretizing a continuous signal into a k-valued signal with $x_{(1)} < \ldots < x_{(k)}$, for $k = 2$ (and $k = 3$ for trinarization) (Fayyad and Irani, 1993; Dougherty *et al.*, 1995; Friedman and Goldszmidt, 1996). The most common approach is to binarize measurements according to a threshold, such as the mean value (Kim *et al.*, 2007), a quantile (Kaiser *et al.*, 2013) or more elaborate estimation techniques as described by Shmulevich and Zhang (2002), Zhou *et al.* (2003) or Hopfensitz *et al.* (2012). A threshold can also be learned from external information (Kestler *et al.*, 2011; Schmid *et al.*, 2014). One commonly assigns the value 0 to the data points below the threshold ('low') and 1 to those above ('high'). The aim of any such binarization method is to find the most appropriate threshold.

Binarization is often part of data preprocessing on which further analyses strongly depend. For example, gene expression data from time series must be binarized for the reconstruction and simulation of Boolean networks (see, e.g., Liang *et al.*, 1998; Lähdesmäki *et al.*, 2003; Müssel *et al.*, 2010). Only then can these networks model gene regulation by representing each gene as a binary variable illustrating the two states 'active' and 'inactive'. In addition to being an important prerequisite to various analyses, binarization and trinarization can also be used independently to extract information. It can, for example, serve as an unsupervised method to reveal patterns

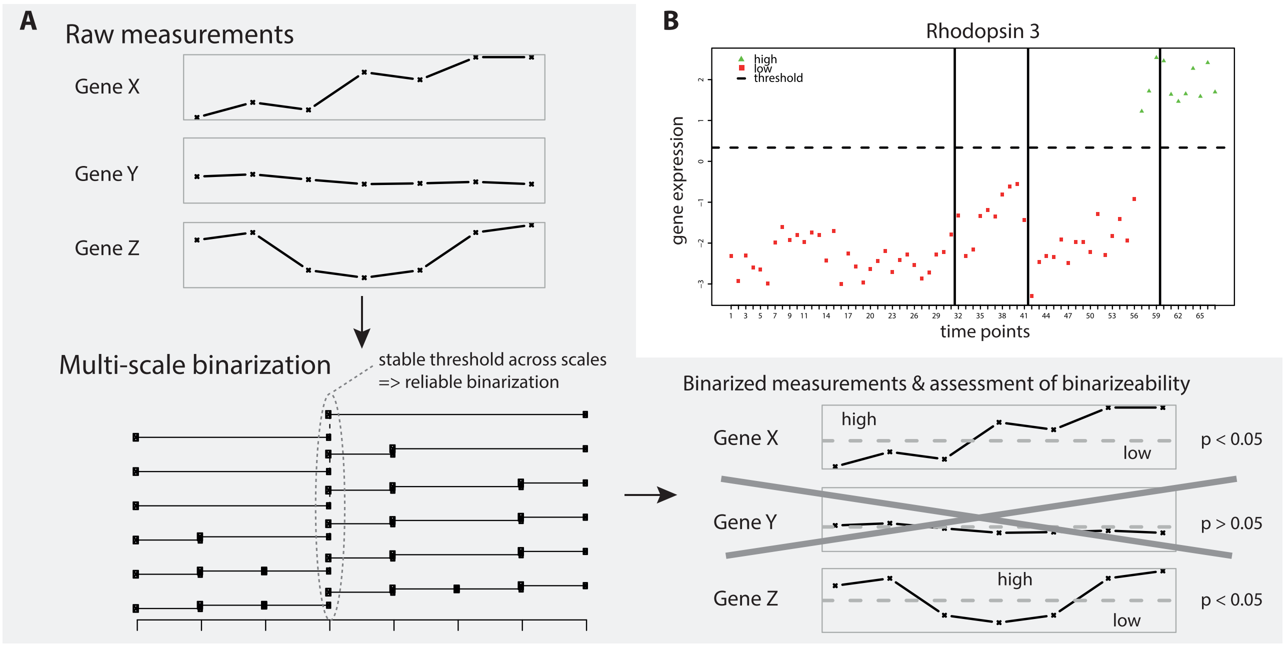

Fig. 1. (**A**) Multiscale binarization. The initial step function representing the measurements for a gene is approximated by a series of smoothed step functions. The strongest discontinuities of these functions are used to calculate a global threshold and to identify unbinarizeable measurements. (**B**) Visualization of a binarization of expression measurements for the gene Rhodopsin 3 using the BASC B algorithm. The horizontal line marks the binarization threshold, whereas the vertical lines separate known major developmental stages of the fruit fly

within data. The benefit of binarized patterns was demonstrated for a variety of diagnostic models for gene expression profiles. Zilliox and Irizarry (2007) showed that binarized signals can be utilized for prototype-based classification. Tuna and Niranjan (2010) demonstrated that the generalization ability of support vector machines is not declined by binarized gene expression profiles.

However, knowledge about data, which cannot be reasonably represented by two groups (i.e. data that is ‘not well binarizeable’) can give valuable hints about the processes that lead to it, like noise or gradients. Nevertheless current implementations of available binarization algorithms (e.g. Kaiser *et al.*, 2013) regardlessly impose an inappropriate separation into two groups and are not able to identify this additional ‘don’t know group’.

We propose the R package `BiTrinA`, which is the first to provide multiscale approaches for binarization, trinarization and reliability assessment.

2 Functionality and application

2.1 Functionality

The `BiTrinA` package implements three binarization and three trinarization algorithms for one-dimensional data with integrated statistical binarizeability and trinarizeability tests and provides visualization functions to analyze the results. All binarization algorithms are based on thresholding a real-valued signal x according to a threshold t

$$f(x) = \mathbb{I}_{[x \geq t]}.$$

Here $\mathbb{I}_{[\cdot]}$ denotes the indicator function.

The first binarization approach is based on k-means clustering (Hartigan, 1975), which heuristically minimizes the sum over the squared distance of the data points to their respective cluster center. It can be seen as a Gaussian mixture model restricted to isotropic local covariance matrices. The k-means algorithm is typically initialized randomly and applied with restarts to avoid suboptimal results. For binarization, the number of clusters k is set to 2. The binarization threshold divides the data points of these two clusters. Hartigan’s dip test for unimodality (Hartigan and Hartigan, 1985; Maechler, 2013) is used to rate the binarizeability of the input data. The sample’s binarizeability (actually its multimodality) is estimated by the distance of the sample distribution F from the class of unimodal distributions U:

$$D(F) = \inf_{G \in U} \sup_{x} |F(x) - G(x)|.$$

Furthermore, multiscale algorithms (BASC A and BASC B; Hopfensitz *et al.*, 2012) are provided. These algorithms compute a series of step functions to obtain a robust binarization threshold (see Fig. 1A). An initial step function is obtained by rearranging the values of the input vector in increasing order. Then, step functions with fewer discontinuities are calculated sequentially. BASC A determines these step functions in such a way that each minimizes the Euclidean distance to the initial step function. Let S_n be the set of all step functions with exactly n discontinuities. An optimal quantization of $f \in S_{N-1}$ by a step function with n steps ($1 \leq n < N - 1$) is defined as

$$s_n^* = \operatorname*{arg\,min}_{s_n \in S_n} ||f - s_n||_2.$$

The distance $||f - s_n||_2$ is the approximation error of a step function s_n with n steps regarding the original function f.

BASC B constructs step functions using a scale space approach. Bessel kernels with different smoothing factors σ are applied to the first derivative of the step function $\Delta(x)$, yielding smoothed derivatives

$$\Delta_\sigma(x) = \sum_{k=-\infty}^{\infty} \frac{\sigma(x)}{e^{2\sigma}} \cdot I_{x-k}(2\sigma).$$

Here, I_n denotes the modified Bessel function of order n. Based on these derivatives, smoothed step functions are constructed and

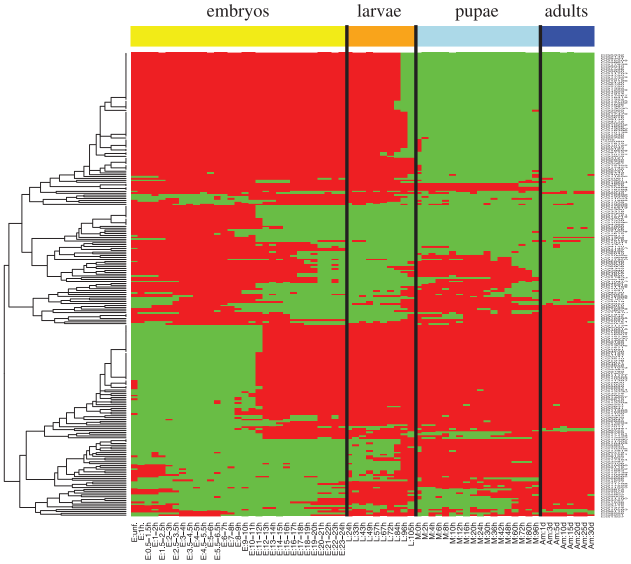

Fig. 2. Heatmap of significantly binarized expression values (BASC B, FDR<0.05) from the Drosophila dataset. Time points are sorted according to developmental stages

selected. For each step function, the strongest discontinuity is determined. For both multiscale algorithms, the binarization threshold is defined as the median of these strongest discontinuities.

A bootstrap test is used to evaluate the robustness of this threshold across the series of functions. A significant *P*-value indicates that the variation of the location of the strongest discontinuity across the step functions is low, i.e. the binarization threshold does not change over multiple scales and the computed binarization is of good quality (Fig. 1A).

For trinarization, the package also provides, apart from *k*-means with $k = 3$, extensions of the BASC algorithms for the Trinarization Across multiple SCales (TASC). The TASC algorithms discretize a continuous signal into three values,

$$f(x) = \mathbb{I}_{[x \geq t_1]} + \mathbb{I}_{[x \geq t_2]},$$

where (t_1, t_2) with $t_1 < t_2$ being a real-valued tuple of thresholds. TASC A or TASC B inherit their optimization strategy from the corresponding BASC algorithms. They select candidate tuples according to their achieved mean discontinuity for each optimal step function. The final thresholds (t_1, t_2) can be seen as the median tuple of all selected candidates. The stability of these thresholds can again be assessed by a bootstrap algorithm.

2.2 Application

The following illustrates the application of the `BiTrinA` package and its statistical tests on gene expression data. We binarized, and trinarized, a dataset containing gene expression measurements (4028 genes, 67 time points) of the developmental stages (embryonic, larval, pupal and adult) of *Drosophila melanogaster* (Arbeitman *et al.*, 2002). We utilized the BASC B algorithm. After correction for multiple testing (FDR < 0.05), 252 of the 4028 genes are statistically significant according to BASC's test for binarizeability. A heatmap of the binarized signals can be found in Figure 2. As an example, Figure 1B illustrates the binarization of the Rhodopsin 3 gene, which is a photoreceptor of the eye. It can be observed that expression remains low throughout the first three developmental stages but then peaks in the adult stage. This is properly reflected by the binarization, which achieves a highly significant *P*-value of $P < 10^{-10}$ for this gene.

Table 1. 10 × 10 crossvalidation results for the *Drosophila melanogaster* data (Arbeitman *et al.*, 2002) (four classes)

Binarization	Accuracy (%) Mean ± SD	Features (%) Mean ± SD	Features (num.) Mean ± SD
—	99.25 ± 0.79	100.00	4028
BASC A	**98.51**±0.00	27.94±0.85	1125.52 ± 34.33
BASC B	97.02 ± 1.00	5.74±0.55	231.17±22.25
k-means ($k = 2$)	89.25 ± 2.88	**0.41**±0.19	16.71 ± 7.49
TASC A	**98.51**±0.00	56.71±0.68	2284.29 ± 27.32
TASC B	**98.51**±0.00	18.18 ± 0.42	732.21 ± 16.81
k-means ($k = 3$)	93.58±1.58	0.41 ± 0.19	16.71 ± 7.49

Note: The mean accuracy of a linear support vector machine (one-versus one, majority vote) and the mean percentage of binarizeable features together with their absolute numbers are shown (FDR < 0.05, best values in bold).

An interesting question with respect to this illustration might be whether the statistical assessment of binarizeability can provide an indication which of the measured genes are involved in the developmental process. That is, can the statistical test identify genes whose binarized measurements somehow reflect the known developmental phases? Statistical evidence indeed suggests that there is a significant association between the concordance of measurements with the developmental stages and their binarizeability (see Supplementary Material). This shows the practical relevance of the proposed quality assessment for the identification of potentially interesting candidate genes for further investigation.

We also performed crossvalidation classification experiments on the binarizeable and trinarizable signals that indicate that indeed the quantized information is sufficient for the categorization of the developmental stages (Table 1). The mean accuracy of a linear support vector machine (Cortes and Vapnik, 1995; Müssel *et al.*, 2012) only declined by about 0.71% for BASC A and the TASC algorithms, and 2.24% for BASC B while utilizing only 27.94% (BASC A), 5.74% (BASC B) and 56.71% (TASC A), 18.18% (TASC B) of all features. In terms of accuracy they outperformed the *k*-means algorithm (10 restarts) that achieved a maximum mean accuracy of 93.58%, with utilizing only 0.41% of all available features.

3 Conclusion

Many analyses require a quantization of data and thus depend heavily on this procedure's results. However, not all data are well binarizeable or trinarizable, and current implementations do not include methods to assess the applicability of the original data for such a method. Our R package `BiTrinA` integrates statistical testing procedures that are capable of assessing the binarizeability and trinarizability, and consequently the quality of discretizing of biological data. As such, the approaches can be used to identify candidate measurements that may be relevant for further analyses.

Funding

The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 602783; the DFG (SFB 1074 project Z1); and the German Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SyStaR, project ID 0315894A) all to H.A.K.

Conflict of Interest: none declared.

References

Arbeitman,M.N. *et al*. (2002) Gene expression during the life cycle of drosophila melanogaster. *Science*, **297**, 2270–2275.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Dougherty,J. *et al*. (1995) Supervised and unsupervised discretization of continuous features. In: Prieditis, A. and Russell, S. (eds) *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, USA.

Fayyad,U. and Irani,K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy, R. (ed.) *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, pp. 1022–1029.

Friedman,N. and Goldszmidt,M. (1996) Discretizing continuous attributes while learning Bayesian networks. In: Saitta, L. (ed.) *ICML*. Morgan Kaufmann, pp. 157–165.

Hartigan,J. (1975) *Clustering Algorithms*. Wiley, New York.

Hartigan,J.A. and Hartigan,P. (1985) The dip test of unimodality. *Ann. Stat.*, **13**, 70–84.

Hopfensitz,M. *et al*. (2012) Multiscale binarization of gene expression data for reconstructing Boolean networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 487–498.

Kaiser,S. *et al*. (2013) *biclust: BiCluster Algorithms. R Package Version 1.0.2.*

Kestler,H. *et al*. (2011) On the fusion of threshold classifiers for categorization and dimensionality reduction. *Comput. Stat.*, **26**, 321–340.

Kim,H. *et al*. (2007) Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*, **8**, 37.

Lähdesmäki,H. *et al*. (2003) On learning gene regulatory networks under the Boolean network model. *Mach. Learn.*, **52**, 147–167.

Liang,S. *et al*. (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **3**, 18–29.

Maechler,M. (2013) *diptest: Hartigan's Dip Test Statistic for Unimodality—Corrected Code. R Package Version 0.75-5.*

Müssel,C. *et al*. (2010) Boolnet—an r package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, **26**, 1378–1380.

Müssel,C. *et al*. (2012) Multi-objective parameter selection for classifiers. *J. Stat. Softw.*, **46**, 1–27.

Schmid,F. *et al*. (2014) Three transductive set covering machines. In: Spiliopoulou, M. *et al* (eds), *Data Analysis, Machine Learning and Knowledge Discovery*. Springer, Heidelberg, pp. 303–311.

Shmulevich,I. and Zhang,W. (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*, **18**, 555–565.

Tuna,S. and Niranjan,M. (2010) Reducing the algorithmic variability in transcriptome-based inference. *Bioinformatics*, **26**, 1185–1191.

Zhou,X. *et al*. (2003) Binarization of microarray data on the basis of a mixture model. *Mol. Cancer Ther.*, **2**, 679–684.

Zilliox,M. and Irizarry,R. (2007) A gene expression barcode for microarray data. *Nat. Methods.*, **4**, 911–913.