 doi:10.20944/preprints202104.0227.v1

Simple method for cutoff point identification in descriptive high-throughput biological studies

Alexander Suvorov

Department of Environmental Health Sciences, School of Public Health and Health Sciences, University of Massachusetts 686 North Pleasant Street Amherst, MA 01003, asuvorov@umass.edu

Abstract

Rapid development of high-throughput omics technologies generates an increasing interests in algorithms for cutoff point identification. Existing cutoff methods and tools identify cutoff points based on association of continuous variables with another variable, such as phenotype, disease state or treatment group. These approaches are not applicable for descriptive studies in which continuous variables are reported without known association with any biologically meaningful variables. The most common shape of the ranked distribution of continuous variables in high-throughput descriptive studies corresponds to a biphasic exponential/super-exponential curve, where the first phase includes big number of variables with values slowly growing with rank and the second phase includes smaller number of variables rapidly growing with rank. This study describes an easy algorithm to identify the boundary between these phases to be used as a cutoff point. The major assumption of that approach is that small number of variables with high values dominate biological system and determine its major processes and functions. This approach was tested on three different datasets: genes in the human cerebral cortex, mammalian genes sensitive to chemical exposures, and proteins expressed in human heart. In every case, the described cutoff identification method produced shortlists of variables (genes, proteins) highly relevant for dominant functions/pathways of the analyzed biological systems. Thus, our described method for cutoff identification may be used to prioritize variables for a focused functional analysis, in situations where other methods of dichotomization of data are inaccessible.

Introduction

Rapid development of high-throughput omics and other similar technologies and approaches in medico-biological domain generates an increasing interests in algorithms for dichotomization of continuous variables. As such, there is a high demand for cutoff point identification tools. Existing methods identify cutoff points based on association of continuous variables with another variable, such as phenotype, disease state or treatment/exposure group. For example, a widely used approach for the identification of genes differentially expressed in relation to a health condition or treatment utilizes fold-change and false-discovery rate adjusted p-value as cutoff criteria. A range of algorithms and online tools was developed to categorize variables for decision making about cancer treatments (Budczies et al., 2012; Camp et al., 2004; Ogłuszka et al., 2019).

Existing approaches for cutoff point identification are not applicable for the identification of the most biologically meaningful results in descriptive studies in which continuous variables (such as gene expression, protein expression, metabolite concentration and similar) are reported without known association with any biologically meaningful variables (phenotype, health condition, treatment group, age etc). For example, high throughput descriptive studies of gene expression in different human

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2021 doi:10.20944/preprints202104.0227.v1

cell/tissue types may provide values of expression of every gene in a tissue. For prioritization of research focus, it may be reasonable to assume that genes with high expression values are more important for the normal tissue physiology than these expression of which is close to zero. However, I was not able to identify any simple method that will allow non-subjective dichotomization of all genes into highly expressed and genes with low level of expression. Similarly, I am not aware of cutoff methods for descriptive data in other omics and similar applications.

The most common shape of the ranked distribution of continuous variables in high-throughput descriptive studies corresponds to biphasic exponential/super-exponential curve (Fig. 1A,C,E), where in a first phase big number of variables have low values. These values increase slowly with the rank. In the second phase, relatively small number of variables demonstrate very rapid growth with their rank number (more rapid then exponential). Thus, the curve of this distribution has a bending point, which delineates the boundary between phases. This boundary may be used as a cutoff point to identify these variables, which dominate in the dataset. However, the complicated shape of the curve makes identification of the bending point a challenging task. Here I present a simple method for the identification of the point in which the curve is changing from exponential to super-exponential growth using Excel or similar graphing tools.

Method description

If we connect the first and the last points of the exponential/super-exponential distribution curve (A) by a straight line (B), together these 2 curves will produce a figure resembling triangle (Fig 1 A,C,E)). Then for every x_A value of the A curve we can calculate a length of a segment that will be perpendicular to the short-cut function (C) (Fig 1A). The longest segment will cross the A distribution curve in its bending point.

The B function is a linear function: $y_B = m_B x_B + b_B$. Functions perpendicular to B, all have the following generic equation: $y_C = (-1/m_B)x_C + b_C$. Given the coordinates of crossing points between A curve and every C function are known (x_{AC} = rank number of the variable, y_{AC} = value of the variable (expression, concentration, abundance, etc)), b_C can be calculated for each such crossing point ($b_C = y_{AC} - (-1/m_B)x_{AC}$). Thus, now for every point of the A curve we have an equation of a linear function C that is passing through it and is perpendicular to the short-cut line B. Now we need to identify coordinates of points at which B and C functions intersect. Given that coordinates of both functions are the same at intersection, we can equate x for both functions: $(y_{CB} - b_B)/m_B = (y_{CB} - b_C)/(-1/m_B)$. From that equation we can calculate y for intersection: $y_{CB} = (b_B + b_C m_B^2)/(1 + m_B^2)$. As we know y for intersection, we can calculate x for intersection as well, using an equation for B: $x_{CB} = (y_{CB} - b_B)/m_B$. Now, as we have coordinates for points of intersection of each C function with A (x_{AC},y_{AC}) and coordinates for intersection of each C function with B (x_{CB},y_{CB}) we can calculate the length of segments using Pythagorean theorem:

$$D = \sqrt{(x_{CB} - x_{AC})^2 + (y_{CB} - y_{AC})^2}.$$

The longest segment will cross A curve in the point of the curve bending.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2021 doi:10.20944/preprints202104.0227.v1

Protocol using MS Excel

1. Generate Excel spreadsheet with three columns: A – name of biological entity (gene, protein, metabolite, DNA site, pathway etc); B – consecutive numbers starting from 1 (x_A); C – values for each variable (expression, concentration, methylation, abundance, enrichment etc) (y_A) sorted from smallest to largest.
2. Find an equation for the line connecting the first and the last points in the ranked distribution. Take coordinates (x,y) of the first and the last points of distribution. Plot them in a separate Excel sheet, add trend line and display equation on chart. This equation will have the following appearance $y_B = m_B x_B + b_B$. Record the same value of m_B and b_B values in every cell of columns D and E respectively in the original spreadsheet.
3. Calculate y-intercept for C functions (b_C) perpendicular to the B function. Insert the following equation in column F to calculate b_C values: $b_C = (-1/m_B)x_{AC} – y_{AC}$. In this equation the value for m_B will be the same for every line, while the values for x_{AC} and y_{AC} will be taken from columns B and C respectively and will be different for every line. These variables have $_{AC}$ index here as they are the same for bot A and C functions in the point of their intersection.
4. Calculate y coordinate for the point of intersection between B and C functions (y_{CB}). Insert the following equation in column G: $y_{CB} = (b_B + b_C m_B^2)/(1 + m_B^2)$. In this equation, values for m_B and b_B will be the same for every line, while the values for b_C will be taken from column F and will be different for every line.
5. Calculate x coordinate for the point of intersection between B and C functions (x_{CB}). Insert the following equation in column H: $x_{CB} = (y_{CB} – b_B)/m_B$. In this equation values for m_B and b_B will be the same for every line, while the values for y_{CB} will be taken from column G and will be different for every line.
6. Calculate the distances (D) of functions C between their intersections with functions A and B. Insert the following equation in the column I:
$$D = \sqrt{(x_{CB} - x_{AC})^2 + (y_{CB} - y_{AC})^2}$$
7. Identify a cutoff point between exponential and super-exponential growth functions. Sort column I values from highest to lowest to identify D with highest value. The ranking number (column B) corresponding to that D value is a cutoff point.

Examples of the Method Use

Example 1: Identification of genes highly expressed in human cerebral cortex

Data on consensus normalized gene expression values in human cerebral cortex were downloaded from The Human Protein Atlas (Uhlen et al., 2015). These values represent the maximum normalized expression values for each gene in three data sources: The Human Protein Atlas, The Genotype-Tissue Expression (GTEx) project (Consortium, 2020) and FANTOM5 (Lizio et al., 2015). The dataset consists of 16,353 genes with their expression values. The distribution of expression values ranked from smallest to largest is shown in figure 1, curve A. The linear function B connecting the first and the last points of the curve A has the following equation: y = 0.0364x – 0.0364. Thus, $m_B = 0.0364$ and $b_B = -0.0364$. Then, b_C can be calculated for each C function perpendicular to B as follows: b_C = (-1/0.0364) x gene rank – gene expression. Coordinates of the intersection point of every C function with B function are calculated as described in previous section (steps 4 and 5). These coordinated were used to calculate the length of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2021 doi:10.20944/preprints202104.0227.v1

segments D using Pythagorean theorem (step 6 in the previous section). The longest segment corresponds to the gene ranked 15,778. This ranking number corresponds to the cutoff point that delineates genes with low and high expression in human cerebral cortex. To test if highly expressed genes reflect essential physiology of cerebral cortex, I submitted the list of top 575 genes, ranking 15,779 through 16,353 to Metascape (Zhou et al., 2019) and conducted enrichment analysis with default settings. The enriched biological categories were highly relevant for the nervous tissue physiology and included for example “nervous system development”, “chemical synaptic transmission”, “cell projection morphogenesis”, “cellular ion homeostasis”, and “learning and memory” among others (Fig. 1B). These categories were enriched with very high level of significance (-log10(p) >15).

Example 2: Identification of genes highly sensitive to chemical exposures

In a recent study sensitivity of genes common to humans, rats and mice was identified based on an overlap of transcriptomic datasets from 2,169 toxicological studies (Suvorov et al., 2021). I use the data from this study available through Mendeley Data (Suvorov et al., 2020). The dataset includes 17,338 genes and their respected chemical sensitivity values, which correspond to the number of individual studies with 1,239 chemical compounds in which gene expression was affected by exposure. Following the same steps as in the previous example, I identified the rank number 15,966 as a cutoff point (Fig. 1C). To test if genes sensitive to chemical exposures are associated with known pathways of response to toxicity, I submitted the list of top 1,373 genes, ranking 15,967 through 17,338 to Metascape. Identified enriched biological categories included many well recognized pathway of response to chemical exposures, stress and damage, for example: “nuclear receptors meta-pathway”, “response to toxic substance”, “apoptotic signaling pathway”, “response to wounding”, “response to oxygen levels”, and “response to oxidative stress” (Fig. 1D). Thus, the cutoff used in this example captured essential molecular mechanisms involved in the response to chemical exposures. These categories were enriched with very high level of significance (-log10(p) >50).

Example 3: Identification of proteins highly expressed in the adult human heart

Data on protein expression (at gene level) in the adult human heart were downloaded from The Human Proteome Map portal (Kim et al., 2014). These data are based on LC-MS/MS utilizing high resolution and high accuracy Fourier transform mass spectrometry. All mass spectrometry data including precursors and HCD-derived fragments were acquired on the Orbitrap mass analyzers in the high-high mode. The dataset includes 17,294 unique gene names and expression values of corresponding proteins. Expression values were calculated as follows: spectral counts per gene per experiment were first summed from all peptides mapped to each gene. Total acquired tandem mass spectra were used to normalize between experiments and then spectral counts per gene were averaged across multiple experiments per tissue. Following the same steps as in previous examples, I identified the rank number 17,086 as a cutoff point (Fig. 1E). To test if proteins shortlisted using my approach reflect essential physiology of the heart, I submitted the list of top 209 genes, ranking 17,087 through 17,294 to Metascape. Top enriched biological categories were highly relevant for the heart physiology and function. These categories included for example “muscle system process”, “oxidation-reduction process” “actin filament-based process”, ‘smooth muscle contraction” and other (Fig. 1F). Thus, the

 doi:10.20944/preprints202104.0227.v1

cutoff used in this example captured essential molecular mechanisms that dominate heart physiology. These categories were enriched with very high level of significance (-log10(p) >10).

Discussion

In this study, I describe a simple and reproducible approach for the cutoff identification in descriptive high-throughput studies, which method is entirely based on the analysis of the shape of the curve of the data distribution. The major assumption of that approach is that small number of variables with high values dominate biological system and determine its major processes and functions. Thus, our described method for cutoff identification may be used to prioritize variables for more detailed functional analysis, in situations where other methods of dichotomization of data are inaccessible.

Three different datasets analyzed here as examples demonstrate that the described cutoff identification method produces shortlists of variables highly relevant for dominant functions/pathways of the analyzed biological systems. In fact, the shortlist of highly expressed genes in the human cerebral cortex was highly enriched for categories related to synaptic transmission, nervous system development and even higher functions, such as learning and memory. The shortlist of genes sensitive to chemical exposures was enriched for biological categories involved in response to stress and damage. Finally, the shortlist of proteins expressed highly in human heart, was significantly enriched for biological categories relevant to muscle architecture, contractions and contraction regulation.

I should note here, that some applications may require more or less stringent criteria for the cutoff. In these situations, our approach may still be useful as it allows to identify the point where the curve of values distribution changes most rapidly. Using this reproducibly identifiable point one may further select criteria with different percent of stringency relative to it. In other words, the cutoff point identified as described here may provide some meaningful reference value. Similarly using of p-value and fold change as cutoff points in omics studies are selected arbitrarily by researchers, but they represent meaningful indicators of the data structure.

The results of the use of the described dichotomization approach should be interpreted cautiously. For example, the fact that some gene was found in the short-list of highly expressed genes in a tissue does not necessarily meant that this gene is highly tissue-specific. In fact many “housekeeping” genes are highly expressed in majority of cell types (Kim et al., 2014), as they are major players of biological processes common for different cells and tissues. It is also likely that some genes with normally low level of expression may still be important players of highly tissue-specific processes. Thus, in each specific situation of the use of the suggested dichotomization approach a biological relevance of the approach should be taken into consideration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2021 doi:10.20944/preprints202104.0227.v1

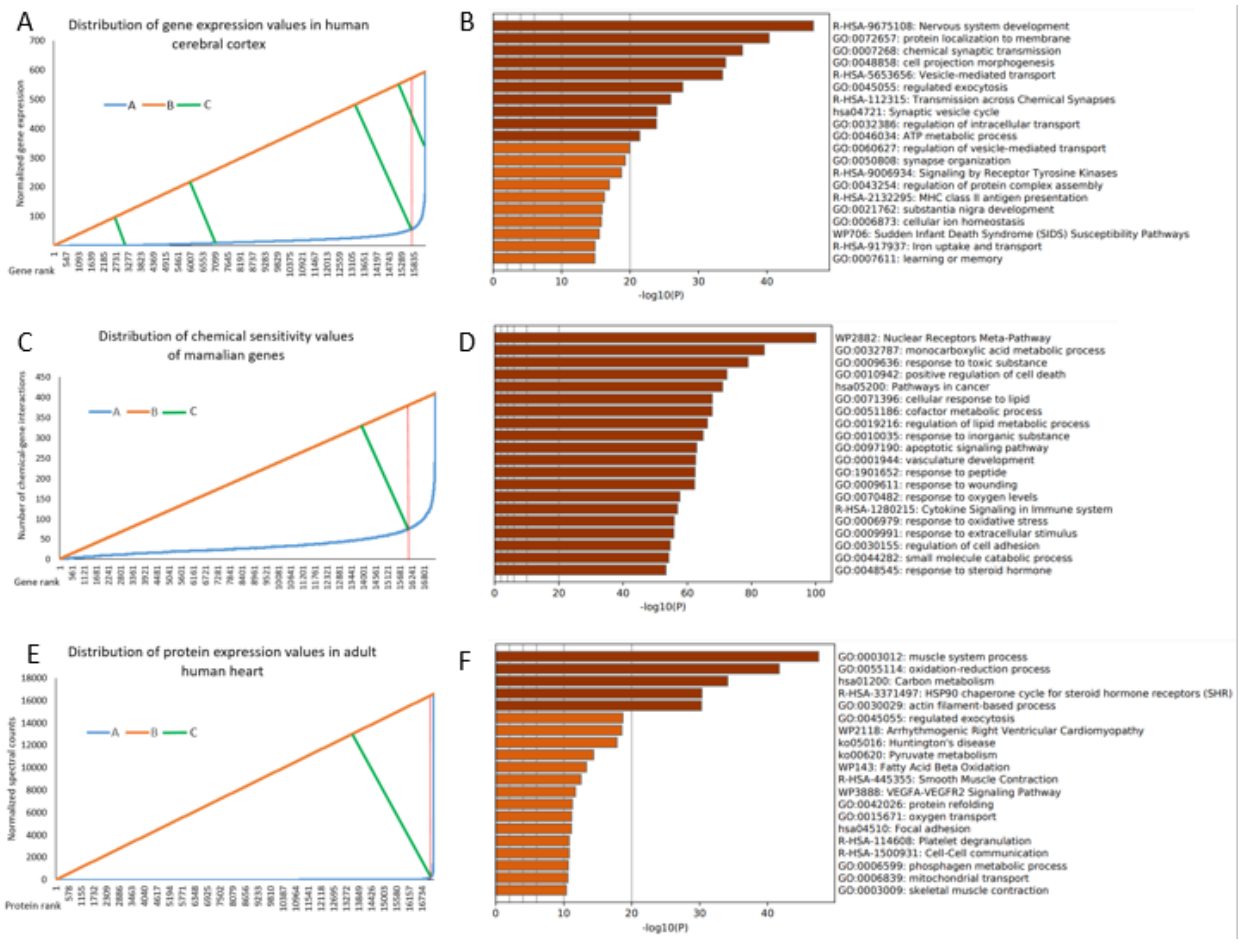

Figure legends

Fig. 1. Illustration of the method for cutoff point identification in descriptive high-throughput biological studies. Variable distribution (A, C, E) and biological categories enriched in shortlists identified using cutoff points (B, D, F) for the following datasets: genes expressed in human cerebral cortex (A, B), genes sensitive to chemical exposures (C, D), and proteins expressed in the adult human heart (E, F). In graphs (A, C, E), A is a curve of the original data distribution, B is a linear short cut connecting the first and the last points of A, and C is a family of linear functions perpendicular to B. Four C functions are shown in figure A. In figures C and D longest segments corresponding C functions are shown. Red vertical lines in figures A,C,E correspond to the cutoff points.

References

Budczies, J., Klauschen, F., Sinn, B.V., Győrffy, B., Schmitt, W.D., Darb-Esfahani, S., Denkert, C., 2012. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PloS One 7, e51862. https://doi.org/10.1371/journal.pone.0051862

Camp, R.L., Dolled-Filhart, M., Rimm, D.L., 2004. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 10, 7252–7259. https://doi.org/10.1158/1078-0432.CCR-04-0713

Consortium, T.Gte., 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. https://doi.org/10.1126/science.aaz1776

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D.N., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S.K., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H., Pandey, A., 2014. A draft map of the human proteome. Nature 509, 575–581. https://doi.org/10.1038/nature13302

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., Mungall, C.J., Arner, E., Baillie, J.K., Bertin, N., Bono, H., de Hoon, M., Diehl, A.D., Dimont, E., Freeman, T.C., Fujieda, K., Hide, W., Kaliyaperumal, R., Katayama, T., Lassmann, T., Meehan, T.F., Nishikata, K., Ono, H., Rehli, M., Sandelin, A., Schultes, E.A., 't Hoen, P.A., Tatum, Z., Thompson, M., Toyoda, T., Wright, D.W., Daub, C.O., Itoh, M., Carninci, P., Hayashizaki, Y., Forrest, A.R., Kawaji, H., FANTOM consortium, 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 16, 22-014-0560–6. https://doi.org/10.1186/s13059-014-0560-6

Ogłuszka, M., Orzechowska, M., Jędroszka, D., Witas, P., Bednarek, A.K., 2019. Evaluate Cutpoints: Adaptable continuous data distribution system for determining survival in Kaplan-Meier estimator. Comput. Methods Programs Biomed. 177, 133–139. https://doi.org/10.1016/j.cmpb.2019.05.023

Suvorov, A., Salemme, V., McGaunn, J., Poluyanoff, A., Amir, S., 2020. Sensitivity of genes, molecular pathways and disease related categories to chemical exposures. Mendeley Data. https://doi.org/10.17632/65fcympd2j.1

Suvorov, A., Salemme, V., McGaunn, J., Poluyanoff, A., Teffera, M., Amir, S., 2021. Unbiased approach for the identification of molecular mechanisms sensitive to chemical exposures. Chemosphere 262, 128362.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., Ponten, F., 2015. Proteomics. Tissue-based map of the human proteome. Science 347, 1260419. https://doi.org/10.1126/science.1260419

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., Chanda, S.K., 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. 10, 1523-019-09234–6. https://doi.org/10.1038/s41467-019-09234-6