Wang et al. BMC Res Notes (2015) 8:832
DOI 10.1186/s13104-015-1786-4

BMC Research Notes

TECHNICAL NOTE

Open Access

EntropyExplorer: an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression

Kai Wang[1], Charles A. Phillips[1*], Arnold M. Saxton[2] and Michael A. Langston[1]

Abstract

Background: Differential Shannon entropy (DSE) and differential coefficient of variation (DCV) are effective metrics for the study of gene expression data. They can serve to augment differential expression (DE), and be applied in numerous settings whenever one seeks to measure differences in variability rather than mere differences in magnitude. A general purpose, easily accessible tool for DSE and DCV would help make these two metrics available to data scientists. Automated p value computations would additionally be useful, and are often easier to interpret than raw test statistic values alone.

Results: *EntropyExplorer* is an R package for calculating DSE, DCV and DE. It also computes corresponding p values for each metric. All features are available through a single R function call. Based on extensive investigations in the literature, the Fligner-Killeen test was chosen to compute DCV p values. No standard method was found to be appropriate for DSE, and so permutation testing is used to calculate DSE p values.

Conclusions: *EntropyExplorer* provides a convenient resource for calculating DSE, DCV, DE and associated p values. The package, along with its source code and reference manual, are freely available from the CRAN public repository at http://cran.r-project.org/web/packages/EntropyExplorer/index.html.

Keywords: Differential Shannon entropy, Differential coefficient of variation, Differential expression, Statistical tests

Background

Shannon entropy (SE) and coefficient of variation (CV) are used to measure the variability or dispersion of numerical data. Such variability has potential utility in numerous application domains, perhaps most notably in the analysis of high throughput biological data. Variability has been applied, for example, to study gene expression data in the context of human disease [1]. Increased entropy in particular, in both gene expression and protein interaction data, has been observed to be a characteristic of cancer [2]. Numerous other examples typify the utility of entropy [3–8] and coefficient of variation [9–12].

Shannon entropy is famously rooted in information theory [13]. To avoid confusion, we emphasize that we use the term "differential entropy" to denote a difference between two Shannon entropy values. This is distinct from information-theoretic terminology, in which "differential entropy" often means the entropy of a continuous, rather than a discrete, random variable [14].

We are particularly interested in differential analysis. In [15], we studied differential Shannon entropy (DSE) and differential coefficient of variation (DCV), and found them highly effective in identifying genes of potential interest not found by differential expression (DE) alone. DSE and DCV are applicable to other types of biological data as well, such as that produced by RNA-Seq technologies, although the usual caveats about careful interpretation apply. The usefulness of DSE and DCV is of course

*Correspondence: cphill25@tennessee.edu
[1] Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996-2250, USA
Full list of author information is available at the end of the article

not limited to biological data. They may be applied to any numerical data for which normalized measures of differential variability are relevant.

Implementation

EntropyExplorer is implemented in R [16]. All features are wrapped into a single function call, which takes as input up to eight arguments. Two of these arguments are numerical matrices, with identical labels for each row. The output is a matrix with two, three or five columns that contains in each row two SE, CV or mean values; a DSE, DCV or DE value; and/or two p values, one raw and one adjusted. Output rows can be sorted by value, raw p value or adjusted p value, and can be filtered to show only the top-ranked rows.

Permutation testing for DSE is accomplished with the help of the R function *sample.int*. The default number of tests to be employed is set to 1000, which the user can override. The p value for DCV is calculated by applying the Fligner-Killeen test for homogeneity of variances, implemented via the R function *fligner.test*, to the log-transform of the input data. The R function *t.test* is used to find a p value for DE. Adjusted p values are calculated using the *p.adjust* function in R, which provides false discovery rate and multiple testing corrections. A more thorough explanation of p value calculations is provided in the discussion section.

EntropyExplorer checks that all matrix entries are positive. This is because calculations of a DSE value/p value and a DCV p value involve taking logarithms, which are undefined on data containing zeros or negative values. Also, CV becomes less meaningful when means approach zero or are negative. Experimental data may be noisy, however, and so *EntropyExplorer* provides mechanisms to handle non-positive values. An optional two-value argument permits the user to add a positive bias to all elements of one or both matrices prior to performing any other calculations. The argument can also be set to make this adjustment automatically, based on the least non-positive value in each matrix.

Metrics

Let $x_1, x_2, \ldots, x_n$ represent a list of n positive numbers, and let $x = \sum_{i=1}^{n} x_i$ denote their sum. The Shannon entropy of this list is

$$SE = \frac{-\sum_{i=1}^{n} \frac{x_i}{x} \log_2 \frac{x_i}{x}}{\log_2 n}.$$

The coefficient of variation is

$$CV = \frac{s}{|\bar{x}|}$$

where $\bar{x} = x/n$ is the sample mean and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ is the sample standard deviation. Given two such lists of positive numbers with Shannon entropies SE_1 and SE_2, coefficients of variation CV_1 and CV_2, and means $\bar{x}_1$ and $\bar{x}_2$, $DSE = |SE_1 - SE_2|$, $DCV = |CV_1 - CV_2|$, and $DE = |\bar{x}_1 - \bar{x}_2|$.

Shannon entropy falls in the range [0, 1]; DSE therefore also falls in the range [0, 1]. Lower (higher) SE corresponds to more (less) variability. CV falls in the range [0, ∞); DCV therefore also has a range of [0, ∞).

Application

EntropyExplorer is invoked as follows:

EntropyExplorer(*expm1, expm2, dmetric, otype, ntop, nperm, shift, padjustmethod*)

We refer the reader to the reference manual, included as Additional file 1 and available on the project webpage, for a detailed description of all arguments and options. Included with the package is a sample mRNA microarray dataset, consisting of a few rows from a dataset obtained from the Gene Expression Omnibus (GEO) [17]. This dataset, GSE10810, contains case/control data on breast cancer [18]. Figures 1 and 2 provide example uses of *EntropyExplorer* on the full data.

Discussion

In addition to calculating DSE, DCV and DE, *EntropyExplorer* can calculate both raw and adjusted p values for each. ANOVA-based tests are the standard way to obtain differential expression p values. We therefore use a t-test for this purpose. Certainly more sophisticated methods exist. See, for example, [19, 20]. Thus, we emphasize that *EntropyExplorer* includes DE only as a simple, convenient and straightforward point of comparison with the other

```
> EntropyExplorer(m1,m2,dmetric="dse",otype="v",ntop=10)
           SE(expm1) SE(expm2) SE(expm1)-SE(expm2)
228245_s_at 0.4585632 0.9749417       -0.5163785
205220_at   0.4347483 0.9494563       -0.5147080
213711_at   0.4658234 0.9705819       -0.5047585
209242_at   0.4757616 0.9597454       -0.4839837
203908_at   0.4971761 0.9460891       -0.4489130
205030_at   0.4183439 0.8599655       -0.4416216
227282_at   0.5773323 0.9901547       -0.4128224
205067_at   0.5584833 0.9406934       -0.3822100
223623_at   0.6033164 0.9801690       -0.3768526
203824_at   0.5562821 0.9249965       -0.3687144
```

Fig. 1 The output of *EntropyExplorer* on breast cancer data. The numerical matrices m1 and m2 have been read into R. The function call has specified "dse" for differential Shannon entropy, "v" for value, and 10 to return the top 10 values

```
> EntropyExplorer(m1,m2,dmetric="dcv",otype="bv",ntop=12)
           CV(expm1) CV(expm2) CV(expm1)-CV(expm2)     p-value  fdr p-value
205220_at   3.970528 0.5873201          3.383208 0.8222579620 0.876318753
213711_at   3.789072 0.5350799          3.253993 0.2336725875 0.354813275
228245_s_at 3.603529 0.4299952          3.173533 0.0118800876 0.046483561
209242_at   3.455212 0.5763780          2.878834 0.0003329751 0.005779234
205067_at   3.399742 0.6848428          2.714899 0.3342106218 0.459651180
207302_at   3.325542 0.8120794          2.513462 0.0314091253 0.087452672
227282_at   2.765104 0.2722402          2.492864 0.1106427072 0.208619781
203908_at   3.066962 0.5786523          2.488309 0.0541907559 0.126061057
226147_s_at 3.530525 1.0483199          2.482205 0.2123513195 0.330828202
223623_at   2.835876 0.3818841          2.453992 0.0001887257 0.004363719
230285_at   3.059179 0.6062872          2.452892 0.0797618163 0.165595404
205752_s_at 2.836653 0.4284138          2.408240 0.7468782641 0.821021185
```

Fig. 2 Another use of *EntropyExplorer* on breast cancer data. The function call has specified "dcv" for differential coefficient of variation, "bv" to specify both value and p value, and to sort by value, and 12 to return the top 12 rows

two metrics. For DCV p values, we observe that 11 tests of equal relative variation were compared in [21], with the conclusion that the Fligner-Killeen test [22] is usually the most appropriate. It strikes a balance between type I and type II errors, and is robust to non-normal distributions.

Obtaining reliable p values for DSE proved much more challenging. We found no known method in the literature specific to DSE p values. We therefore investigated the extent to which SE is correlated to variance. A high correlation would suggest that they may be proxies for each other, in which case the p value of an F-test or some derivation thereof might serve as suitable estimate of the DSE p value. Unfortunately, correlations between SE and variance, or between SE and a function of variance, were not high enough to justify using one as a surrogate for the other. Table 1 shows the correlation between SE and variance V, and between SE and the function $\frac{1}{2}\ln(2\pi eV)$ as an attempt to linearize the relationship, using the 16 datasets from [15]. The only notably high correlation is found in the obesity dataset. The obesity data, however, contains a large number of missing values, rendering the high correlation less reliable. We conclude that standard statistical tests related to variance do not appear suitable for testing DSE.

We also examined the distribution of DSE on the 16 datasets, with the goal of empirically determining a suitable reference distribution for DSE. From this, we could then estimate p values analytically. We applied the Kolmogorov–Smirnov (KS) test to compare the DSE distribution of each dataset to some of the more common reference distributions, such as normal, F, t, and Chi square. When performing a KS test, p values can be overly sensitive to deviations from the reference distribution [23], so a D-statistic value below 0.1 was used to identify matching distributions. In our experiments, only the Parkinson's dataset produced a D-statistic below 0.1 when tested against a normal or standardized t distribution (Table 2). Figure 3 shows a sample distribution of DSE, in this case using prostate cancer data.

Table 1 Correlations between SE and variance, and between SE and $\frac{1}{2}\ln(2\pi eV)$, on 16 microarray gene expression datasets

Datasets	Correlation Between SE and Variance		Correlation between SE and $\frac{1}{2}\ln(2\pi eV)$	
	Case	Control	Case	Control
Allergic Rhinitis	−0.5515	−0.5769	−0.9703	−0.9658
Asthma_GSE4302	−0.4272	−0.4677	−0.1924	−0.2004
BreastCancer_GSE10810	−0.3942	−0.3378	−0.1810	−0.1265
CLL_GSE8835	0.2251	0.2522	−0.0806	−0.0624
ColorectalCancer_GSE9348	0.3122	0.4454	−0.0086	0.0206
CrohnsDisease_GSE6731	−0.2826	−0.2380	−0.1664	−0.4020
LungAdenocarcinoma_GSE7670	0.0725	0.3360	−0.0173	0.0105
MS_GDS3920	−0.3615	−0.3320	−0.0515	−0.0559
Obesity_GSE12050	0.9998	0.9990	0.1584	0.5420
Pancreas_GDS4102	−0.4137	−0.4455	−0.1331	−0.0890
ParkinsonsDisease_GSE20141	−0.1732	−0.2554	−0.0024	−0.0155
ProstateCancer_GSE6919_GPL8300	0.2118	0.1552	−0.0562	−0.0699
Psoriasis_GSE13355	−0.6386	−0.6554	−0.5200	−0.6779
Schizophrenia_GSE17612	0.3632	0.3910	0.0170	0.0235
T2D_GSE20966	−0.6006	−0.5550	−0.4356	−0.4663
UlcerativeColitis_GSE6731	−0.3112	−0.2555	−0.1799	−0.1451

Table 2 KS test D-statistic results comparing the DSE distribution against several common distributions

Dataset	Distribution				
	Normal	Chi-square	F	t	t (standardized DSE)*
Allergic Rhinitis	0.3109	1	1	0.4991	0.3526
Asthma_GSE4302	0.2795	1	1	0.4895	0.3117
BreastCancer_GSE10810	0.2115	1	1	0.4797	0.3944
CLL_GSE8835	0.1506	1	0.9975	0.4519	0.1596
ColorectalCancer_GSE9348	0.1232	1	0.9994	0.4514	0.2142
CrohnsDisease_GSE6731	0.2131	1	0.987	0.4691	0.2392
LungAdenocarcinoma_GSE7670	0.19	1	0.9999	0.4663	0.332
MS_GDS3920	0.2703	1	0.9994	0.4813	0.3397
Obesity_GSE12050	0.2352	1	0.9991	0.484	0.287
Pancreas_GDS4102	0.2606	1	0.9937	0.4532	0.3254
ParkinsonsDisease_GSE20141	0.0628	1	0.9361	0.3816	0.0582
ProstateCancer_GSE6919_GPL8300	0.1575	1	1	0.4739	0.2522
Psoriasis_GSE13355	0.3327	1	0.9999	0.4932	0.4195
Schizophrenia_GSE17612	0.183	1	0.9998	0.4705	0.2138
T2D_GSE20966	0.3271	1	0.9999	0.4936	0.3562
UlcerativeColitis_GSE6731	0.2397	1	0.998	0.4831	0.3608

* The last column shows the results after first standardizing DSE by dividing each DSE by the standard deviation of all DSEs

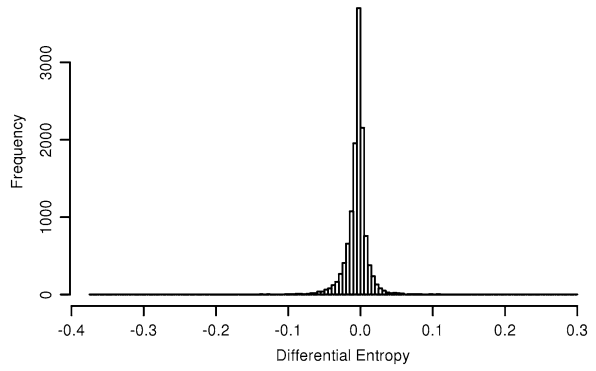

Fig. 3 The distribution of differential Shannon entropy. The observed distribution of differential Shannon entropy in sample prostate cancer data is shown. Similar patterns were seen in all 16 data sets. None of the standard distributions tested matched the observed distributions closely enough to be considered as a reference distribution for obtaining p values

We conclude from this that none of the distributions tested are close enough approximations to the observed DSE distribution to be used as a proxy for obtaining p values. Thus, without a known distribution function or suitable surrogate, we resort to resampling in order to obtain reliable DSE p values. While computationally demanding, the following permutation test makes no assumptions about the underlying distribution of the data. Given two lists of numbers, containing n_1 and n_2 numerical elements respectively, we first calculate their DSE and then create a new list A containing all $n_1 + n_2$ numbers from the two lists. Next we randomly permute the elements of A, then recalculate DSE, treating the first n_1 elements of A as one list and the last n_2 elements of A as a second list. The resultant p value is simply the proportion of all recalculated DSEs that are at least as extreme as the original DSE.

In addition to raw p values, *EntropyExplorer* also calculates p values adjusted for multiple testing. A user can choose to adjust based on FDR, Holm or another multiple-testing adjustment.

Conclusions

We have produced *EntropyExplorer*, an R package for calculating differential Shannon entropy, differential coefficient of variation and differential expression. This package also calculates raw and adjusted p values for each metric. These measures have been shown to complement one another [15], making this package an effective tool for users in search of more expansive suites of differential analysis methods.

Availability and requirements

Project name: *EntropyExplorer.*

Project home page: http://cran.r-project.org/web/packages/EntropyExplorer/index.html.

Operating system(s): Platform independent.

Programming language: R.

Other requirements: R version 3.0 or later is recommended.

Any restrictions to use by non-academics: None.

Additional availability: *EntropyExplorer* is integrated into the GrAPPA toolkit at http://grappa.eecs.utk.edu/.

Additional file

Additional file 1. EntropyExplorer R package reference manual.

Authors' contributions

KW implemented the package and performed numerous analytical tests. CAP led exhaustive software evaluations and isolated relevant data. AMS provided statistical expertise and assisted with human factors engineering. MAL directed the project and provided for its support. All authors participated in writing the paper. All authors read and approved the final manuscript.

Author details

[1] Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996-2250, USA. [2] Department of Animal Science, University of Tennessee Institute of Agriculture, Knoxville, TN 37996-4574, USA.

Acknowledgements

This work has been supported in part by the National Institutes of Health under awards R01-AA-018776 and 3P20MD000516-07S1.

Competing interests

The authors declare that they have no competing interests.

Received: 12 August 2015 Accepted: 2 December 2015
Published online: 30 December 2015

References

1. Ho JW, Stefani M, dos Remedios CG, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. Bioinformatics. 2008;24(13):i390–8.
2. West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Scientific reports. 2012;2:802.
3. Berretta R, Moscato P. Cancer biomarker discovery: the entropic hallmark. PLoS One. 2010;5(8):e12262.
4. Sherwin WB. Entropy and information approaches to genetic diversity and its expression: genomic geography. Entropy. 2010;12(7):1765–98.
5. Masisi L, Nelwamondo F, Marwala T. The use of entropy to measure structural diversity. In: IEEE 6th International Conference on Computational Cybernetics. Stará Lesná, Slovakia.
6. Chen B-S, Li C-W. On the interplay between entropy and robustness of gene regulatory networks. Entropy. 2010;12(5):1071–101.
7. Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinform. 2003;4:54.
8. Kohane IS, Kho AT, Butte AJ. Microarrays for an integrative genomics. Cambridge, Mass.: MIT Press; 2003.
9. Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. Clin Diagn Lab Immunol. 2002;9(6):1235–9.
10. Weber EU, Shafir S, Blais AR. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychol Rev. 2004;111(2):430–45.
11. Faber DS, Korn H. Applicability of the coefficient of variation method for analyzing synaptic plasticity. Biophys J. 1991;60(5):1288–94.
12. Bedeian AG, Mossholder KW. On the use of the coefficient of variation as a measure of diversity. Organ Res Methods. 2000;3(3):285–97.
13. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):45.
14. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken: Wiley-Interscience; 2006.
15. Wang K, Phillips CA, Rogers GL, Barrenas F, Benson M, Langston MA. Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes. Int J Comput Biol Drug Des. 2014;7(2–3):183–94.
16. R Core Team. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing. Vienna, Austria; 2014.
17. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
18. Pedraza V, Gomez-Capilla JA, Escaramis G, Gomez C, Torne P, Rivera JM, Gil A, Araque P, Olea N, Estivill X, et al. Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. Cancer. 2010;116(2):486–96.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
20. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 2001;98(9):5116–21.
21. Donnelly SM, Kramer A. Testing for multiple species in fossil samples: an evaluation and comparison of tests for equal relative variation. Am J Phys Anthropol. 1999;108(4):507–29.
22. Fligner MA, Killeen TJ. Distribution-free two-sample tests for scale. J Am Stat Assoc. 1976;71(353):210–3.
23. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10(2):486–9.