BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

CatchMore: an R package for richness estimation

Qijin Kendrick Li & Amy D Willis [1,*]

[1] Department of Biostatistics, University of Washington.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The vast genetic diversity of microorganisms inevitably results in undersampling. Estimating the diversity of genomes and genes is a challenging statistical problem and existing informatics tools can be unstable or difficult to use.

Results: We present a R package to estimate the number of unobserved microbial genomes or genes in a population based on a sample. The method fits a large number of zero-truncated mixed- and non-mixed Poisson models to the observed frequency counts and then performs model selection to return a best estimate of total diversity. A convenient interface is provided to allow testing for differences in total diversity.

Availability: The R package CatchMore is freely available from CRAN and github.

Contact: adwillis@uw.edu

1 INTRODUCTION

The genetic diversity of a microbial community, either at the gene- or genome-level, is a common indicator of microbial or functional health in both host-associated and environmental microbiology. However, the total diversity of an environment is rarely known, and typically needs to be estimated based on a sample of the environment. Despite extensive statistical literature concerning the estimation of diversity, there exists limited accessible and stable software to estimate diversity. CatchAll (**?**), a historically tool, implements mixed-Poisson models for estimating total diversity. However, CatchAll's source code is not publicly available and its command line executable requires an emulator to run on Unix systems.

We have created an open-source R package that combines the stability of CatchAll with modern statistical models for estimating species richness, and more user-friendly interface. Our software is called CatchMore.

To illustrate CatchMore, we estimate the number of total human host associated genomes based on data from **?**.

(Note that while we use "species" the parameter of interest is actually defined by whatever the unit of dataentry is, i.e., we could be estimating gene diversity.)

2 THEORY & METHOD

Let f_j be the number of species observed j times and $C = \sum_{j=0}^{\infty} f_j$ be the total number of species in the population. We can calculate $f_1, f_2, \ldots$ but need to predict f_0, the number of species observed zero times. The estimate of total richness is $\hat{C} = \hat{f}_0 + \sum_{j \geq 1} f_j$.

Table 1. Comparison of CatchMore with CatchAll and breakaway... Think about what needs to go in here: runtime, estimates

Dataset	CatchMore	breakaway	CatchAll	
	83.41	210.95	169.61	263.53
0%	83.39	5.00	4.49	5.98
-80%	83.70	158.05	136.40	162.90

Extrapolating from $f_1, f_2, \ldots$ to predict f_0 is numerically and statistically unstable, and so statistical estimation of C typically involves a generative model for the frequency counts, such as mixed-Poisson with specified mixing distribution.

CatchMore includes the following estimates:

1. Poisson distributed
2. Mixed-Poisson with exponential mixing
3. Mixed-Poisson with mixture of two exponentials
4. Mixed-Poisson with mixture of three exponentials
5. ...

3 RESULTS & DISCUSSION

Algorithmic details and ... are available as Supplementary Material.

4 CONCLUSION

The key contribution of CatchMore is a statistically stable, easily accessible and open source software tool for estimating diversity in microbiome and other datasets...

ACKNOWLEDGEMENT

The authors would like to thank John Bunge and Linda Woodard for providing the CatchAll source code...

Funding: None declared.

[*] to whom correspondence should be addressed

REFERENCES

Bunge, J., Böhning, D., Allen, H., and Foster, J.A. (2012) Estimating population diversity with unreliable low frequency counts, *Pac Symp Biocomput*, **12**, 203–212.

Bunge, J., Woodard, L., Böhning, D., Foster, J.A., Connolly, S., and Allen, H.K. (2012) Estimating population diversity with CatchAll, *Bioinformatics*, **28** (7), 1045–1047.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population, *Scandinavian Journal of statistics*, 265–270.

Willis, A.D., and Bunge, J. (2015) Estimating Diversity via Frequency Ratios, *Biometrics*.