BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 7 2012, pages 1045–1047
doi:10.1093/bioinformatics/bts075

Genetics and population analysis

Advance Access publication February 13, 2012

Estimating population diversity with CatchAll

John Bunge[1,*], Linda Woodard[2], Dankmar Böhning[3], James A. Foster[4], Sean Connolly[5] and Heather K. Allen[6]

[1]Department of Statistical Science, [2]Center for Advanced Computing, Cornell University, Ithaca, NY 14853, USA, [3]School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK, [4]Department of Biological Sciences, University of Idaho, Moscow, ID 83844, [5]Charles River Associates, Boston, MA 02116 and [6]Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, Agricultural Research Service, Ames, IA, 50010, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: The massive data produced by next-generation sequencing require advanced statistical tools. We address estimating the total diversity or *species richness* in a population. To date, only relatively simple methods have been implemented in available software. There is a need for software employing modern, computationally intensive statistical analyses including error, goodness-of-fit and robustness assessments.

Results: We present CatchAll, a fast, easy-to-use, platform-independent program that computes maximum likelihood estimates for finite-mixture models, weighted linear regression-based analyses and coverage-based non-parametric methods, along with outlier diagnostics. Given sample 'frequency count' data, CatchAll computes 12 different diversity estimates and applies a model-selection algorithm. CatchAll also derives discounted diversity estimates to adjust for possibly uncertain low-frequency counts. It is accompanied by an Excel-based graphics program.

Availability: Free executable downloads for Linux, Windows and Mac OS, with manual and source code, at www.northeastern.edu/catchall.

Contact: jab18@cornell.edu

Received on December 1, 2011; revised on February 4, 2012; accepted on February 6, 2012

1 INTRODUCTION

The field of microbial ecology is bursting with data from next-generation sequencing, but analysis remains a challenge. Estimating the diversity of a microbial community is especially important. To model this statistically, assume that the population can be divided into a finite number of classes. The simplest definition of *diversity* is the number of classes C. A sample drawn from such a population will typically have repeated observations of the various classes: some may be observed once only, others twice and so on, while many classes may not appear in the sample at all. The 'frequency count' data is $\{(i,f(i)), i \geq 1\}$ where $f(i)$ is the number of sample classes of size i. For example, the dataset $\{(1,10),(2,4),(3,2),(7,1)\}$ has 10 'singletons', four 'doubletons', … and one class occurring seven times in the sample. For bacterial and phage diversity, the counts are derived from the frequencies of 16S rRNA genes and contig spectra.

Statistical estimation of diversity, from frequency count data goes back to 1943 (Bunge and Barger, 2008), but so far only coverage-based non-parametric methods have been implemented in widely available software, because they do not require numerical optimization. We improve upon previous methods by (i) implementing parametric finite-mixture models and a new weighted linear regression approach in addition to existing non-parametric methods; (ii) providing a way to statistically discount large numbers of potentially artifactual rare species; and (iii) applying our analyses to highly diverse phage metagenomes.

2 METHODS

We introduce CatchAll version 3.0. [A preliminary version, lacking key capabilities presented here, was discussed in (Bunge, 2011).] The program computes 12 different diversity estimates with standard errors and goodness-of-fit assessments, at every level of outlier deletion. It proposes a best overall parametric estimate along with a ranked set of alternatives. For cases where low-frequency counts may be erroneous, CatchAll computes a discounted estimate by adjusting the highest diversity component of the selected mixture model. CatchAll is fast, platform-independent, computationally robust, and has both batch and GUI interfaces. An associated Excel spreadsheet automatically produces graphical displays.

CatchAll computes three types of analyses. (i) Finite mixture models (Bunge and Barger, 2008). A convex combination of distributions is fitted to the observed count data, yielding a diversity estimate, standard error and goodness-of-fit statistics. Five models are computed: order 0 $\equiv$ Poisson; orders 1–4 $\equiv$ mixtures of 1–4 geometric distributions. Maximum likelihood estimation is done via a nested double expectation–maximization (EM) algorithm. (ii) Weighted linear regression model (Rocchetti *et al.*, 2011). We fit a linear regression model to $(i, r(i) := (i+1)f(i+1)/f(i))$. The ratio r is a linear function of i under the Poisson and negative binomial models, and can be robust to departures from these. Inherent heteroscedasticity requires weighted regression. (iii) Coverage-based estimates (Chao and Lee, 1992). These are based on non-parametric adjustments to the sample *coverage* $\equiv$ the proportion of the population represented in the sample. CatchAll computes Good-Turing and Chao1 as lower bounds; the Abundance-Based Coverage Estimator (ACE) and its high-diversity variant ACE1; and Chao-Bunge, which is optimal under the negative binomial model.

Exceptionally abundant classes tend to generate high sample frequencies, which can lead to poor model fit or unstable estimates. As a check, we delete every point above some maximum frequency τ; we then compute every analysis at every τ. For the parametric models, a selection algorithm combines χ^2 goodness-of-fit tests, AIC and other criteria, to select an optimal model and cutoff τ: essentially the 'best' selected model admits the largest τ while maintaining acceptable AIC- and χ^2-based goodness-of-fit. For the

*To whom correspondence should be addressed.

WLRM, we select between log-transformed and untransformed versions, and choose maximum feasible τ. For the non-parametric methods, either ACE or ACE1 is chosen according to the coefficient of variation of the data based on published criteria, at $\tau \leq 10$. (See user manual for full details.) The best selected analyses, along with close alternatives, and analyses computed at maximum τ, are presented in the GUI and in a 'Best Models Analysis' file. Complete information is given in 'Analysis' and 'Fits' files.

Our selection algorithms provide choices within families of models (parametric, weighted linear regression, coverage-based non-parametric), but do not address choice between families. The user may regard the selected results for the parametric, weighted linear regression and non-parametric methods ('Best', 'WLRM', and 'NonP 2' in Table 1) as comparable *grosso modo*, although their underlying statistical assumptions differ considerably. The final choice of method is at the discretion of the user.

In some cases, the sample low-frequency counts may be questionable; for instance, when the counts are based on potentially erroneous DNA sequence matching (Behnke *et al.*, 2011). In order to statistically reduce the importance of the low-abundance species in such cases, the best fitted mixture model is computed and its highest diversity component, i.e. the component of the mixture model representing a smoothed version of a proportion of the lowest frequency counts, is deleted. This yields a discounted total diversity estimate (Bunge *et al.*, 2012), which is reported in the GUI and the Best Models Analysis file.

3 EXAMPLE

Phage diversity analyses represent a new level of population diversity beyond what is encountered in other areas of microbial ecology. We illustrate the application of CatchAll to a contig spectrum from a swine fecal metagenome (Allen *et al.*, 2011). The contig spectrum was generated using Circonspect via the CAMERA pipeline (Sun *et al.*, 2011). The complete dataset is [(1,4736), (2,521), (3,152), (4,69), (5,46), (6,27), (7,21), (8,18), (9,16), (10,10), (11,9), (12,8), (13,7), (14,6), (15,5), (16,4), (17,4), (18,3), (19,3), (20,3), (21,3), (22,2), (23,2), (24,3), (25,3), (26,1), (27,2), (28,1), (29,2), (30,2), (31,1), (32,1), (33,1), (34,1), (35,1), (36,1), (37,1), (38,1), (39,1), (40,1), (41,1), (42,0), (43,1), (44,0), (45,1), (46,0), (47,0), (48,0), (49,0), (50,0), (51,0), (52,1)]. CatchAll output (slightly abbreviated here) as displayed in the GUI screen or equivalently in the 'Best Models Analysis' file is shown in Table 1.

This analysis took 309s in GUI mode on a 3 GHz/8 MB RAM 64 bit notebook PC. Computation time depends on the complexity (in particular, the smoothness) of the frequency count data not the original sample size, because the original sequence data are reduced to frequency counts before analysis.

In this case, the best fitted parametric model and its first two alternatives (2a and 2b) are the same, and the third alternative (2c) is very close. The various analyses agree approximately at optimal τ, with Chao1 serving as a lower bound, while some anomalies are seen at max τ, as expected; in particular, ACE and ACE1 should only be used for $\tau \leq\approx 10$, the value of Non-P $\tau_{\max}$ is displayed only for comparative purposes.

CatchAll selects the the log-transformed version of the weighted linear regression model at $\tau = 5$, still agreeing with the other analyses albeit with a larger SE. This demonstrates the robustness of the WLRM, since it is theoretically optimal for data with lower diversity than our phage example.

The best discounted model steps down from a three- to a two-component mixture, and reduces the estimated total diversity by 97.4%, from 67 792 (SE 8656) to 1727 (SE 221). At present, there is no formal statistical hypothesis test to select the original versus the discounted models, so the choice depends on the investigator's level of confidence in the low-frequency counts. This is a topic of current research.

Table 1. CatchAll analysis of phage metagenomic diversity data

Obs = 5703	Model	τ	Est Div	SE	Lwr CB	Upr CB
Best	3Mixed	52	67 792	8656	53 009	87 195
2a	3Mixed	52	67 792	8656	53 009	87 195
2b	3Mixed	52	67 792	8656	53 009	87 195
2c	2Mixed	10	64 683	5473	54 893	76 421
WLRM	LogTrans	5	63 103	13 352	42 306	95 718
NonP 1	Chao1	2	27 229	1141	25 106	29 584
NonP 2	ACE1	10	68 790	4620	60 365	78 514
Parm $\tau_{\max}$	3Mixed	52	67 792	8656	53 009	87 195
WLRM $\tau_{\max}$	LogTrans	41	22 107	2535	17 842	27 870
Non-P $\tau_{\max}$	ACE1	52	422 854	55 507	327 457	546 534
Best Disc	2Mixed	52	1727	221	1410	2305

Obs, observed number of species; Est Div, estimated total diversity; SE, standard error; Lwr CB, Upr CB, lower and upper 95% confidence bounds (respectively). Best, 2a, 2b, 2c, top four selected parametric models; WLRM, weighted linear regression model; NonP 1, Chao1; NonP 2, ACE or ACE1 as selected; Parm $\tau_{\max}$, WLRM $\tau_{\max}$, Non-P $\tau_{\max}$, given models at max τ. See program manual for details.

ACKNOWLEDGEMENTS

We thank Rob Knight, Tom Casey, Uri Levine and Thad Stanton for helpful discussions; and Sam Humphrey for technical support.

Funding: Development of CatchAll was funded by National Science Foundation grant (DEB-08-16638) to J.B. This research was conducted using the resources of the Cornell Center for Advanced Computing, which receives funding from Cornell University, the National Science Foundation, and other leading public agencies, foundations and corporations. J.A.F was funded in part by NIH COBRE P20RR16448, NIH INBRE P20RR016454, and NSF STC DBI-0939454 'BEACON Center for the Study of Evolution in Action'. The conclusions and views expressed herein are the conclusions and views of the authors and do not reflect or represent the views of Charles River Associates, which accepts no duty of care or liability of any kind whatsoever to any party in connection with this artical. USDA is an equal opportunity employer.

Conflict of Interest: none declared.

REFERENCES

Allen,H.K. *et al.* (2011) Antibiotics in feed induce prophages in swine fecal microbiomes. *mBio*, **2**; doi: 10.1128/mBio.00260-11.

Behnke,A. *et al.* (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Envrion. Microbiol.*, **13**, 340–349.

Bunge,J. (2011) Estimating the number of species with CatchAll. In *Biocomputing 2011: Proceedings of the Pacific Symposium, Kohala Coast, Hawaii, USA, 3-7 January 2011*. World Scientific Publishing, Hackensack, New Jersey, USA.

Bunge,J. and Barger,J. (2008) Parametric models for estimating the number of classes. *Biometr. J.*, **50**, 971–982.

Bunge,J. *et al.* (2012) Estimating population diversity with unreliable low frequency counts. In: *Biocomputing 2012: Proceedings of the Pacific Symposium, Kohala Coast, Hawaii, USA, 2-6 January 2012*. World Scientific Publishing, Hackensack, New Jersey, USA.

Chao,A. and Lee,S.M. (1992) Estimating the number of classes via sample coverage. *J. Am. Stat. Associ.*, **87**, 210–217.

Rocchetti,I. *et al.* (2011) Population size estimation based upon ratios of recapture probabilities. *Ann. Appl. Stat.*, **5**, 1512–1533. doi: 10.1214/10-AOAS436.

Sun,S. *et al.* (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.*, **39**, D546–D551.