Bioinformation by Biomedical Informatics Publishing *open access*

www.bioinformation.net **Web Server**

RICHEST - a web server for richness estimation in biological data

Chris Durden and Qunfeng Dong*

Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana, USA; Qunfeng Dong – E-mail: dongq@indiana.edu; *Corresponding author

received January 22, 2009; accepted February 05, 2009; published February 27, 2009

Abstract:
Richness is defined as the number of distinct species or classes in a sample or population. Although richness estimation is an important practice, it requires mathematical and computational methods that are challenging to understand and implement. We have developed a web server, RICHness ESTimator (RICHEST), which implements three non-parametric statistical methods for richness estimation. Its user-friendly web interface allows users to analyze and compare their data conveniently over the web.

Availability: A web server hosting RICHEST is accessible at http://richest.cgb.indiana.edu/cgi-bin/index.cgi and the software is freely available for local installations.

Keywords: diversity estimation; biological data; complex; simple; population

Background:

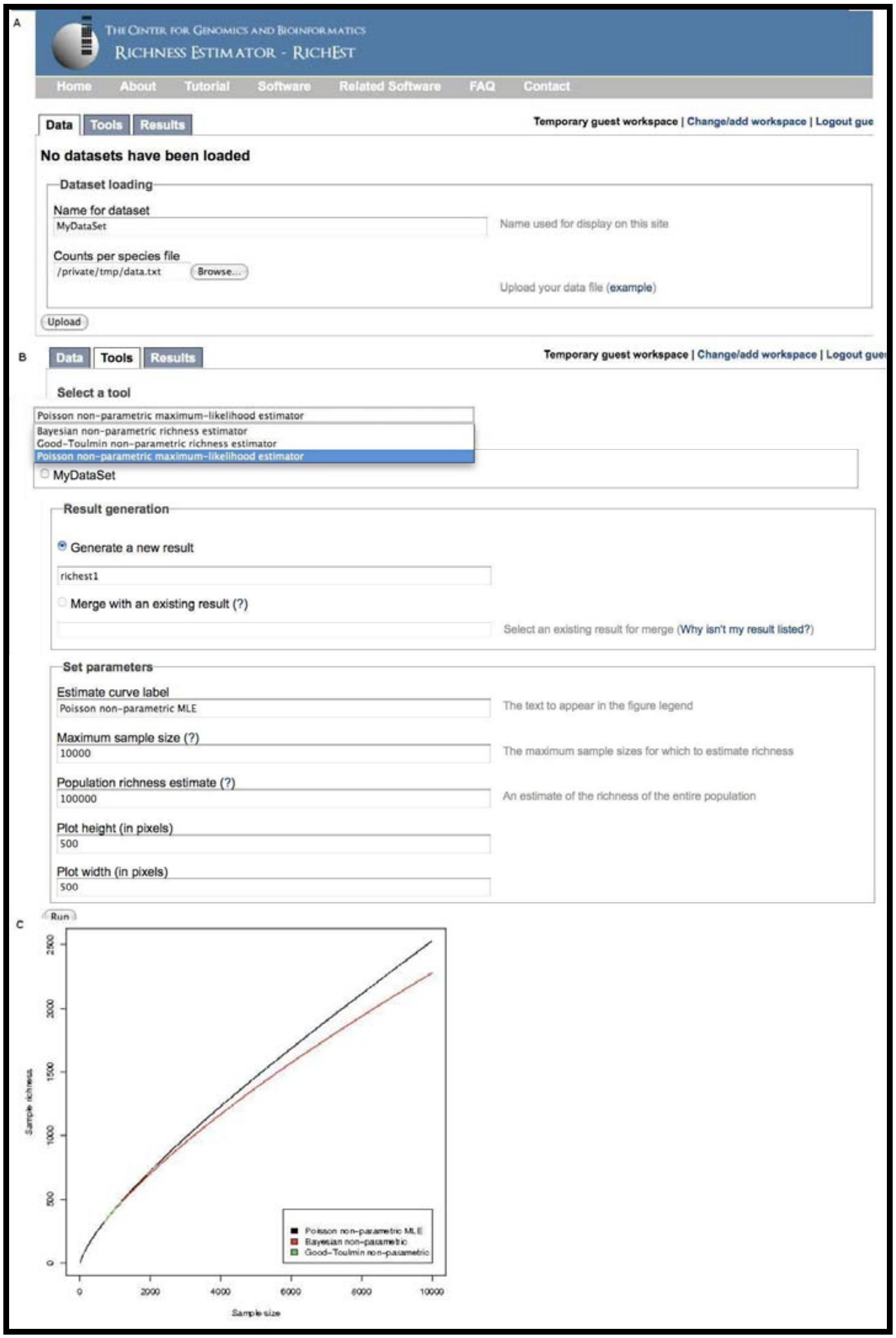
Characterizing biological diversity has been a long-standing interest among biologists. In particular, estimating species richness (i.e., the number of distinct plant, animal or microbial species) is important for understanding the structure of specific environmental niches. Because it is impractical to exhaustively survey entire ecosystems, such estimates must often be inferred from sampling, e.g., given the number of unique species identified from an initial random sample, how many new species are expected in an additional sample? It is not trivial to perform such estimations. First, it is still an active research area for statisticians to develop robust estimation methods. Biologists do not always have time to follow the latest statistical breakthroughs. Second, methods published by statisticians are not always implemented into convenient computer programs for biologists. Biologists are left with mathematical formulations that may be prohibitively complicated to use with their data. We have developed a web-based server, RICHness ESTimator (RICHEST), which implements sophisticated statistical methods for richness estimation. The intuitive web interface guides users to upload their data and easily perform their analyses.

Methodology:

Implementation:

We have identified three non-parametric richness estimation methods: (i) a non-parametric Bayesian method described by Lijoi et al. **[1]**; (ii) a Poisson non-parametric maximum-likelihood (PNPML) method, similar to Wang et al. **[2]**; paired with the rarefaction methodology **[3]**; and (iii) the Good-Toumlin estimator **[4]**. We have chosen to focus on non-parametric methods because they eliminate the need for assumptions about the underlying distribution. However, despite the elegant mathematics published in the Bayesian method, no computer program accompanies the publication. The computational instructions obtained directly from the authors (personal communication) are difficult to follow without significant computational expertise. For the PNPML method, Wang et al. **[2]** have already implemented it into a user-friendly web-based tool called ESTstat. However, the focus of the ESTstat web server is for analyzing EST data although users can also apply it for general species richness estimation. Other existing tools for richness estimation such as EstimateS **[5]** are desktop applications instead of web-based servers.

To bridge the gap between the published statistical methods and the biologists wishing to apply the methods to their data, we have implemented all three methods in RICHEST using the R program for statistical computing. For the Bayesian method, the estimates are achieved by optimizing Pitman's sampling formula with respect to two parameters of a Poisson-Dirichlet distribution. In our implementation, we have taken advantage of the smoothness of the function of Pitman's sampling formula to apply the Nelder-Mead optimization technique **[6]**. The implementation of the PNPML method is similar to that of ESTstat **[2]**, except that it does not implement a penalty function to constrain the population richness parameter. Instead, a user can specify some estimate of the richness of the population and use the method to estimate the richness of subsequent samples. When such information is not scientifically available, one strategy for using the PNPML method is to set the population richness estimate parameter to the Chao lower bound, which is automatically calculated by RICHEST when the data is loaded. The PNPML method uses an expectation-maximization algorithm **[7]** to estimate the maximum-likelihood species abundance distribution based on the non-parametric Poisson mixture model, and it uses this estimated distribution to generate sample richness estimates by rarefaction. Finally, the classic Good-Toulmin estimator was implemented as described in **[4]**. The web interface of RICHEST **(Figure 1)** is implemented with the Perl programming language.


A
THE CENTER FOR GENOMICS AND BIOINFORMATICS
RICHNESS ESTIMATOR - RICHEST
Home
About
Tutorial
Software
Related Software
FAQ
Contact
Data
Tools
Results
Temporary guest workspace | Change/add workspace | Logout gue
Browse...
Upload
B
Data
Tools
Results
Temporary guest workspace | Change/add workspace | Logout gue
Run
C
Sample richness
0
500
1000
1500
2000
2500
Sample size
0
2000
4000
6000
8000
10000
■ Poisson non-parametric MLE
■ Bayesian non-parametric
■ Good-Toulmin non-parametric

No datasets have been loaded

Dataset loading

Name for dataset: MyDataSet — Name used for display on this site

Counts per species file: /private/tmp/data.txt — Upload your data file (example)

Select a tool

Poisson non-parametric maximum-likelihood estimator

- Bayesian non-parametric richness estimator
- Good-Toulmin non-parametric richness estimator
- Poisson non-parametric maximum-likelihood estimator

○ MyDataSet

Result generation

● Generate a new result: richest1

○ Merge with an existing result (?) — Select an existing result for merge (Why isn't my result listed?)

Set parameters

Estimate curve label: Poisson non-parametric MLE — The text to appear in the figure legend

Maximum sample size (?): 10000 — The maximum sample sizes for which to estimate richness

Population richness estimate (?): 100000 — An estimate of the richness of the entire population

Plot height (in pixels): 500

Plot width (in pixels): 500

Figure 1: Screenshots of the RICHEST program outlining key features of the application. The RICHEST web interface consists of three tabs, named Data, Tools, and Results that divide the analysis process into three stages as described below. Users are encouraged to follow the tutorial at the project web site to demonstrate the step-by-step usage of the program. (A) Loading data: The Data tab allows the user to load tab-delimited data files. Users may also upload multiple datasets before running any estimation procedures. (B) Selecting tools: The tools tab allows the user to select and run our integrated richness estimation programs. The tab prompts the user to select which data to use as input, which method to use for estimation, and which options to use for that method. (C) Viewing results: RICHEST outputs a hyperlink to a tab-delimited table and the corresponding graph representing the estimated sample richness curve which gives the richness estimates as a function of the cumulative sample size.

Validation:
We have validated the RICHEST tool by applying it to several data sets. Here, we describe a published microbial data set compiled by Schloss and Handelsman **[8]**, in which total 690 16S rRNA sequences were extracted from the Sargasso Sea metagenome sequencing project **[9]** and clustered into 114 different microbial species based on sequence similarity. To validate the methods that we have integrated in RICHEST, we used a subset of the 16S rRNA sequences to predict richness of the whole microbial sample. For example, we randomly selected some 16S rRNA sequences and counted the number of microbial species in this subset; then we used the three methods to estimate the total number of predicted microbial species in all 690 16S rRNA sequences. That is, we asked whether our predictions would be close to what Schloss and Handelsman observed (i.e., 114 microbial species in total). Table 1 (under supplementary material) shows that our predictions are consistent with their observations. Our result also confirms the prediction by Lijoi et al that the Good-Toulmin estimator provides reliable estimates for the species richness of an additional sample only if its size is not too large.

Acknowledgements:
This work was supported in part by the Indiana METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. We would like to thank Drs. Esfandiar Haghverdi and Zhijun Wu for their helpful discussion on the RICHEST development. We also thank Chris Hemmerich for his assistance in the web server development. We are grateful to Dr. Rameses Mena for providing instructions on the implementation of the Bayesian method.

References

[1] A. Lijoi *et al., BMC Bioinformatics,* (2007) 8:339 [PMID:17868445]
[2] J. P. Wang *et al., BMC Bioinformatics,* (2005) 6:300 [PMID:16351717]
[3] J. C. Tipper, *Paleobiology,* (1979) 5:423.
[4] I. J. Good & G.H. Toulmin, *Biometrika,* (1956) 43:45.
[5] http://purl.oclc.org/estimates
[6] J. A. Nelder & R. Mead, *Comput J.* (1965) 7:308
[7] Z. Liu et al., *Comp Stat & Data Anal* (2006) 50:1052
[8] P. D. Schloss *et al., Appl Environ Microbiol.* (2005) 71:1501 [PMID:15746353]
[9] J. C. Venter *et al., Science* (2004) 304:66 [PMID:15001713]

Edited by P. Kangueane
Citation: Durden & Dong, Bioinformation 3(7): 296-298 (2009)

Supplementary material

Table 1: Predictions of the RICHEST programs on 16S rRNA sequence data

N	Bayesian	PNPML	Good-Toumlin
400 (86)	114.7 (6.1)	113.0 (6.4)	114.1 (7.3)
500 (97)	114.2 (4.2)	108.2 (4.7)	114.0 (4.6)
600 (106)	114.0 (2.8)	112.5 (3.2)	114.0 (2.8)

The first column provides the size of initial sampling of the 16S rRNA accompanied by the observed number of microbial species in parenthesis. The remaining columns provide microbial species richness predictions for the 690 16S rRNA sequences in the study from Bayesian, PNPML, and Good-Toumlin methods, respectively. The predictions are accompanied by standard deviations in parenthesis computed from repeating the experiment 10 times. The total number of observed species (i.e., the truth) is 114 [8].