BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 7 2009, pages 964–966
doi:10.1093/bioinformatics/btp021

Genome analysis

QVALITY: non-parametric estimation of q-values and posterior error probabilities

Lukas Käll[1,2,*], John D. Storey[3] and William Stafford Noble[2,4]

[1]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, Sweden, [2]Department of Genome Sciences, University of Washington, Seattle, WA, [3]Lewis-Sigler Institute, Princeton University, Princeton, NJ and [4]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Received on October 24, 2008; revised on December 23, 2008; accepted on January 7, 2009

Advance Access publication February 4, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: QVALITY is a C++ program for estimating two types of standard statistical confidence measures: the q-value, which is an analog of the p-value that incorporates multiple testing correction, and the posterior error probability (PEP, also known as the local false discovery rate), which corresponds to the probability that a given observation is drawn from the null distribution. In computing q-values, QVALITY employs a standard bootstrap procedure to estimate the prior probability of a score being from the null distribution; for PEP estimation, QVALITY relies upon non-parametric logistic regression. Relative to other tools for estimating statistical confidence measures, QVALITY is unique in its ability to estimate both types of scores directly from a null distribution, without requiring the user to calculate p-values.

Availability: A web server, C++ source code and binaries are available under MIT license at http://noble.gs.washington.edu/proj/qvality

Contact: lukas.kall@cbr.su.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A common feature of high-throughput experiments is that they generate large amounts of data of variable quality. Such data require associated statistical confidence measures. A common way to derive such a measure is by comparing the observed score distribution with the distribution generated by a model representing the noise of the process, a so called *null model*. The null model can either be empirical—i.e. the analysis is repeated in a setting that permutes the data or the labels—or, if we have sufficient knowledge about the noise in the process, we can derive an analytical model.

Given these two distributions, the experimenter must decide what statistical confidence measure to use. This decision depends on whether they want to draw conclusions regarding a set of data points or they are interested in characterizing individual data points. The QVALITY software calculates two complementary and widely used statistical confidence measures: the *q-value* and the *posterior error probability* (PEP). To understand these two measures, consider a set of observations ranked according to scores $s_1,\ldots,s_n$. The p-value of a score s_i is defined as the probability of observing a score as extreme as or more extreme than s_i, assuming that the null hypothesis is correct. The q-value is an analog of the p-value that incorporates multiple testing correction (Storey and Tibshirani, 2003). Specifically, the q-value associated with score s_i is defined as the minimal false discovery rate (FDR) level at which s_i would be deemed significant. Thus, although the q-value is associated with a single observation, it is fundamentally a rate and hence is a property of the collection of scores $s_1,\ldots,s_i$. On the other hand, the PEP of score s_i is simply the probability that this score is drawn according to the null hypothesis. In the statistics literature, PEP is sometimes referred to as the local false discovery rate.

Given a set of p-values, computing corresponding false discovery rates and hence q-values is relatively straightforward (Storey and Tibshirani, 2003); however, computing accurate PEPs is considerably more difficult. Indeed, given accurate PEPs, computing the FDR is trivial: the FDR is simply the sum of PEPs of the significant examples divided by the number of significant examples (Storey *et al*., 2005). The converse is not true. Deriving PEPs from FDRs would lead to an estimated PEP with high variance. In QVALITY, we instead build upon a previously described non-parametric regression method (Anderson and Blair, 1982), modifying it for PEP estimation. Similar methods have been used previously in the analysis of microarray gene expression data (Efron *et al*., 2001; Storey *et al*., 2005).

The QVALITY software is a C++ standalone executable that calculates q-values and PEPs. Many existing tools can be used to estimate one or both of these confidence measures [reviewed in Strimmer (2008)]. However, unlike other tools, QVALITY does not require that the user provide p-values or z-scores as input; instead, the user may simply input two sets of scores: the observed distribution and an empirical null distribution. The q-values are estimated directly (Storey, 2002), and the PEPs are estimated using non-parametric logistic regression. Significantly, QVALITY does not fit two individual distributions for the alternative and the null hypotheses, but models the ratio between the two distributions. QVALITY provides both a command line and a library interface, and a web server is available for users who do not want to download the software.

[*]To whom correspondence should be addressed.

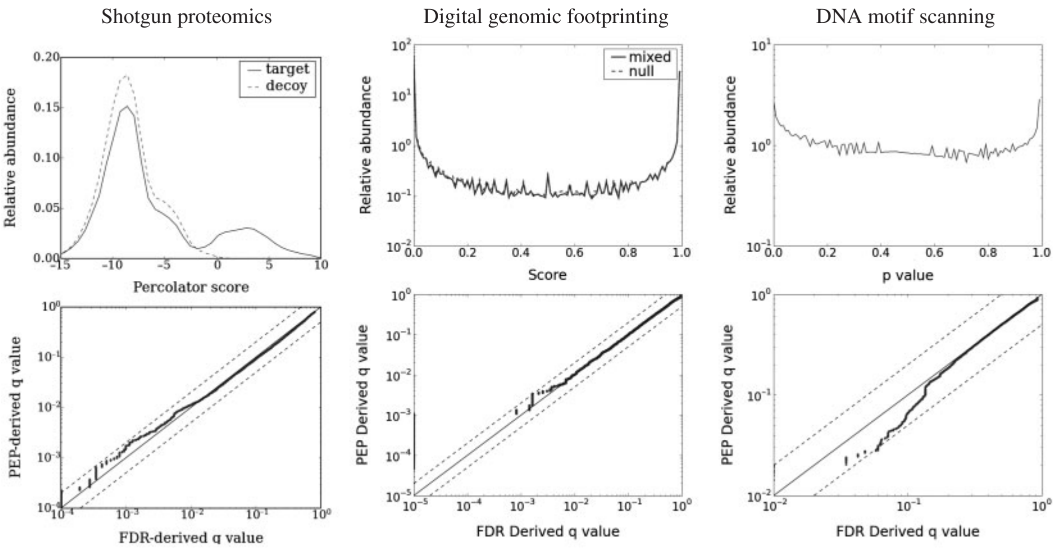

Fig. 1. Application of QVALITY to three different datasets. Each panel in the top row plots the observed score distribution and (in two cases) the corresponding empirical null distribution. The three applications are shotgun proteomics, digital genomic footprinting and DNA motif scanning. The second row of panels illustrates the accuracy of the inferred q-values, plotting the quantiles of the distributions of q-values estimated directly from the empirical null and indirectly via the PEP estimates. In the bottom three panels, the dotted lines correspond to the lines $y = 2x$ and $y = 0.5x$.

2 EXAMPLES OF QVALITY FUNCTIONALITY

In Figure 1, we demonstrate the broad utility of QVALITY by applying it to three diverse bioinformatics applications. Details of each dataset are given in the online supplement. Briefly, the three applications are as follows.

First, we used shotgun proteomics to generate fragmentation spectra, and we assigned a peptide to each spectrum using a database search procedure coupled with a machine learning post-processor. In the top left panel of Figure 1, the series labeled ‘target’ is the distribution of observed scores, and the ‘decoy’ distribution is generated by searching against a database of shuffled sequences. In this example, the hypothesis we want to test is whether we have successfully identified the peptide that generated the observed spectrum.

The bottom left panel of Figure 1 demonstrates the accuracy of the PEPs estimated by QVALITY. The figure plots the quantiles of two different q-value distributions, one computed directly from the empirical null following the methodology of Storey (2002), and the other computed indirectly by first computing PEPs and then integrating. The latter method is expected to be less accurate, because we can measure the false discovery rate directly from the observed distributions of scores, while the PEPs are calculated using a smoothed estimate of the ratio between the observed and empirical null distribution. The figure shows that the two sets of q-values agree within a factor of two over four orders of magnitude.

The second application involves estimating statistical confidence scores for protein-binding footprints observed in a DNaseI-based cleavage assay. Here, the empirical null is derived by locally shuffling the cleavage counts. Again, Figure 1 shows the empirical and null distributions (top middle panel) and the accuracy of the estimated q-values (bottom middle panel).

QVALITY can also be applied to data for which analytical p-values are available. To illustrate this functionality, we scanned the ENCODE regions of the human genome with a position-specific scoring matrix representing the binding affinity of the DNA-binding protein CTCF. The right-hand panels in Figure 1 show the empirical distribution of these p-values, as well as the accuracy of the inferred q-values.

3 USAGE AND IMPLEMENTATION

QVALITY takes two files as input, containing the empirical and null score distributions. If the null distribution is not provided, then the empirical scores are interpreted as p-values. The program produces a three-column file listing the raw score, the q-value and the PEP as output. Note that, when a user applies an empirical null model, it is preferable to provide the observed and null scores separately, rather than precomputing p-values for input to QVALITY. This is because QVALITY requires the null scores for the PEP estimation procedure. If only p-values are provided, then QVALITY must estimate the null scores themselves from the p-values.

The input data are pooled together and binned into 500 equally sized bins. The fraction of null scores is calculated for each bin separately. A set of 2D spline knots are constructed from the median scores of the bins and the fractions of null scores. Thereafter, an interpolating natural cubic spline (Green and Silverman, 1994) is fitted to the spline knots with an iteratively reweighted least squares technique, and the roughness penalty is set to minimize cross-validation error using golden section search. Our approach is similar to that of Efron *et al*. (2001) and is described in more detail in Käll *et al*. (2008).

Computing q-values and PEPs for 70 000 scores and the same number of null scores takes ∼3 s on a 2.33 GHz Intel Xeon processor. The computation time and memory usage scales linearly with the size of the input.

Funding: This work was funded by NIH awards R01 EB007057 and P41 RR11823.

Conflict of Interest: none declared.

REFERENCES

Anderson,J.A. and Blair,V. (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, **69** 123–136.

Efron,B. *et al*. (2001) Empirical bayes analysis of a microarray experiment. *J. Am Stat. Assoc.*, **96**, 1151–1161.

Green,P.J. and Silverman,B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC, Boca Raton, FL.

Käll,L. *et al* (2008) Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*, **24**, i42–i48.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc.*, **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Storey,J.D. *et al*. (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.*, **3**, 1380–1390.

Strimmer,K. (2008) A unified approach to false discovery rate estimation. *BMC Bioinformatics*, **9**, 303.