BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1

Supplement to "QVALITY: Nonparametric estimation of q values and posterior error probabilities"

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

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

Following are descriptions of the three applications shown in Figure 1.

1 EXAMPLE 1: THE PEPTIDE IDENTIFICATION PROBLEM IN SHOTGUN PROTEOMICS

Shotgun proteomics aims to identify the proteins in a complex sample. The process involves cleaving the proteins with an enzyme, separating the resulting peptides by liquid chromatography and detecting the peptides with a tandem mass spectrometer. The peptide identification problem entails mapping an observed spectrum back to the peptide that generated it. This problem is most often solved by searching the fragmentation spectrum against a database of theoretical spectra derived from the the amino acid sequences of the organism being studied. Each peptide-spectrum match (PSM) is given a score by the search engine. A variety of machine learning algorithms have been used to improve the score function and to calibrate top-scoring PSMs across multiple spectra. In the top left panel in Figure 1, the series labeled "target" is the distribution of scores generated by the semisupervised learning algorithm Percolator [Käll et al., 2007] with respect to a collection of spectra from a yeast whole-cell lysate. The "decoy" distribution is our empirical null distribution. This distribution is generated by searching the same spectra against a database of shuffled peptide sequences. In this example, the hypothesis we want to test is whether the PSM is correct or not; i.e., whether we have successfully identified the peptide that generated the observed spectrum.

2 EXAMPLE 2: DIGITAL GENOMIC FOOTPRINTING

The second application of QVALITY involves estimating statistical confidence scores for protein-binding footprints observed in a DNaseI-based cleavage assay [Hesselberth et al., 2008]. Briefly, the digital genomic footprinting assay involves lightly digesting a nuclear DNA sample with DNaseI, size selecting short fragments, and sequencing the ends of these fragments. With sufficient sequencing coverage, individual protein binding footprints can be ascertained as regions of depleted DNaseI cleavage against a background of high DNaseI cleavage. We use a hypergeometric score function to rank candidate footprints, and a greedy selection procedure to eliminate overlapping candidates. An empirical null score distribution is generated by repeating the entire procedure on a locally shuffled version of the data.

3 EXAMPLE 3: DNA MOTIF SCANNING

In the first two examples, we inferred confidence measures by using an emipirical null model. However, QVALITY can also be applied to data for which analytical p values are available. To illustrate this functionality, we used FIMO (`http://meme.sdsc.edu`) to scan the ENCODE regions of the human genome with a position specific scoring matrix representing the binding affinity of the DNA-binding protein CTCF [Kim et al., 2007]. FIMO computes p values using a standard dynamic programming procedure [Staden, 1994].

REFERENCES

J. Hesselberth, Z. Zhang, X. Chen, P. J. Sabo, R. Sandstrom, R. E. Thurman, M. Weaver, S. Neph, M. S. Kuehn, M. O. Dorschner, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. Submitted, 2008.

L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–25, 2007.

T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245, 2007.

R. Staden. Searching for motifs in nucleic acid sequences. *Methods in Molecular Biology*, 25:93–102, 1994.