BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 11 2011, pages 1569–1570
doi:10.1093/bioinformatics/btr165

Sequence analysis
Advance Access publication April 5, 2011

DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models

Cenny Taslim[1,2], Tim Huang[1] and Shili Lin[2,*]

[1]Department of Molecular Virology, Immunology and Medical Genetics and [2]Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: Differential Identification using Mixtures Ensemble (DIME) is a package for identification of biologically significant differential binding sites between two conditions using ChIP-seq data. It considers a collection of finite mixture models combined with a false discovery rate (FDR) criterion to find statistically significant regions. This leads to a more reliable assessment of differential binding sites based on a statistical approach. In addition to ChIP-seq, DIME is also applicable to data from other high-throughput platforms.

Availability and implementation: DIME is implemented as an R-package, which is available at http://www.stat.osu.edu/~statgen/SOFTWARE/DIME. It may also be downloaded from http://cran.r-project.org/web/packages/DIME/.

Contact: shili@stat.osu.edu

Received on January 19, 2011; revised on March 11, 2011; accepted on March 28, 2011

1 INTRODUCTION

ChIP-seq, chromatin immunoprecipitation (ChIP) assay followed by sequencing technology, is gaining popularity as the experiment to analyze genome-wide protein–DNA interactions at high resolution. ChIP-seq experiment produces millions of short sequences that needs to be aligned to a reference genome. The location where a significant number of short reads overlapped on the genome (sometimes referred to as a peak) has been shown to coincide with chromatin immunoprecipitation enrichment, indicating the binding site of the protein of interest.

A number of methods have been proposed to identify peaks in ChIP-seq data, such as FindPeaks, MACS and CisGenome. For more details on these and other existing algorithm including their comparisons, see Laajala *et al.* (2009). These softwares focus on identifying the peaks in one sample or in comparison with a matching input DNA. Whereas in DIME, we focus on identifying differential binding sites of a specific protein under two different biological conditions. The input to DIME is normalized differences of ChIP-seq counts, as in our recent work (Taslim *et al.*, 2009). Specifically, in that paper, we proposed a method to normalize and classify the enrichment regions that are significantly different between two ChIP-seq samples (Taslim *et al.*, 2009). Since then we have thoroughly compared the fitting of ours with another mixture model in the literature (Dean and Raftery, 2005), proposed an ensemble approach to synthesize advantages from different approaches and developed an R-package (DIME) to implement this method. By ensemble, we mean using a collection of three classes of mixture models where the best overall model is selected using BIC and AIC criteria (Hastie *et al.*, 2009). Based on the best overall model, we classify each observation using local FDR (Khalili *et al.*, 2009). One class of models used is the Normal-Uniform model (NUDGE) (Dean and Raftery, 2005). We have modified this class to improve the fit by allowing for multiple normals, leading to the class of iNUDGE models. The third class of models considered is the Gamma-Normal-Gamma (GNG) mixture (Khalili *et al.*, 2009), which was adopted in our recent ChIP-seq analysis (Taslim *et al.*, 2009). In GNG, the differential sites are captured by two exponential components as opposed to a uniform distribution as in iNUDGE. Further, for GNG and iNUDGE, some of the normal components may represent differential sites as well. As such, the applicability of DIME is greatly extended beyond uniform or exponential, since any distribution can be well approximated by a mixture of normals. By utilizing multiple models in a single analysis, we are able to improve the fit of the model to the data, which in turn improves the performance of the differential analysis for data from various omic platforms.

2 DESCRIPTION OF R FUNCTIONS

The main function of the package is `DIME`, which performs model fitting followed by classification of differential binding sites using normalized differences of ChIP-seq counts. Below we briefly describe its usage and functions.

Input: the only required input is an R list that contains normalized differences for chromosome(s) that need to be analyzed. Each element of the list contains data from one chromosome. Users can conveniently include/exclude chromosome(s) to/from the list depending on which chromosome(s) are of interest. The following would be an example of a call to the main function DIME with some optional parameters:

```
result <- DIME(data,gng.tol = 1e − 5,gng.max.iter = 2000,
            gng.K = 2,gng.fdr.cutoff = 0.1),
```

where data is an R list as described above.

Optional input parameters are available to control the fit and classification process. In the example above, two convergence criteria can be specified by users: `gng.max.iter` assigns the maximum number of iterations for fitting GNG model (default = 2000) and `gng.tol` (default = 1e-5) specifies the L^2 norm of differences in the GNG parameter estimates in the current and previous iterations. Thus, the algorithm will stop whenever either

[*]To whom correspondence should be addressed.

one of these two criteria is satisfied. The maximum number of normal components searched when fitting GNG model is set using `gng.K` (default = 2). User can also adjust the local FDR for GNG-based classification by setting `gng.fdr.cutoff` (default = 0.1). Similar optional input parameters are available for the other two classes of models. Other additional optional input parameters are discussed in details in the manual of the R package.

Output: the output of `DIME` is a list containing four elements each corresponding to the results of the overall best model or those of an individual class. The results include details about the fit and classification, such as the estimated parameters for each component of the mixture and mixing proportions. For example, if the best overall model is GNG, then `result$best = result$gng`. Thus, `result$best$pi` gives the estimated proportion of each components under the GNG model. Further, `result$best$mu` and `result$best$sigma` provide the estimated means and SDs, respectively, for the normal components under GNG. A complete list of all available parameters and their descriptions are described in the user guide.

Our package also provides a number of graphical functions that produce plots for the best model as well as for the three individual classes. In the next section, we will give examples of these functions which were used to produce the figures displayed here.

3 EXAMPLE

To demonstrate the utility of the package, DIME was run to compare ChIP-seq data of a normal breast cancer cell line (MCF7) before and after Estradiol (E2) treatment with Polymerase II antibody (Taslim *et al.*, 2009). The models were estimated using the default parameters, except for the maximum number of normal components which was set to be 5. Thus, the command used to run the analysis was as follows: `result <−DIME(data, gng.K = 5, inudge.K = 5)`, where data are the normalized difference of ChIP-seq counts before and after E2 treatment (dataset included in the R package). It took 4675 s on an AMD quad-core 2.4 GHz processor to fit all three classes of models (searching up to five normal components) on around 20k genes (data points) using one random seed. As with any other random search algorithm, the running time of DIME is dependent on how good the initial parameters are. Figure 1A depicts the QQ-plot of the observed data against iNUDGE, GNG and NUDGE (from left to right), which are generated using `inudge.plot.qq`, `gng.plot.qq`, and `nudge.plot.qq`, respectively. The program selected GNG as the best overall model as it provides the best fit compared with iNUDGE and NUDGE, which is evident in Figure 1(A). Figure 1(B) shows the GNG density plot along with its individual components superimposed on the histogram of the normalized data, generated using the `gng.plot.fit` function.

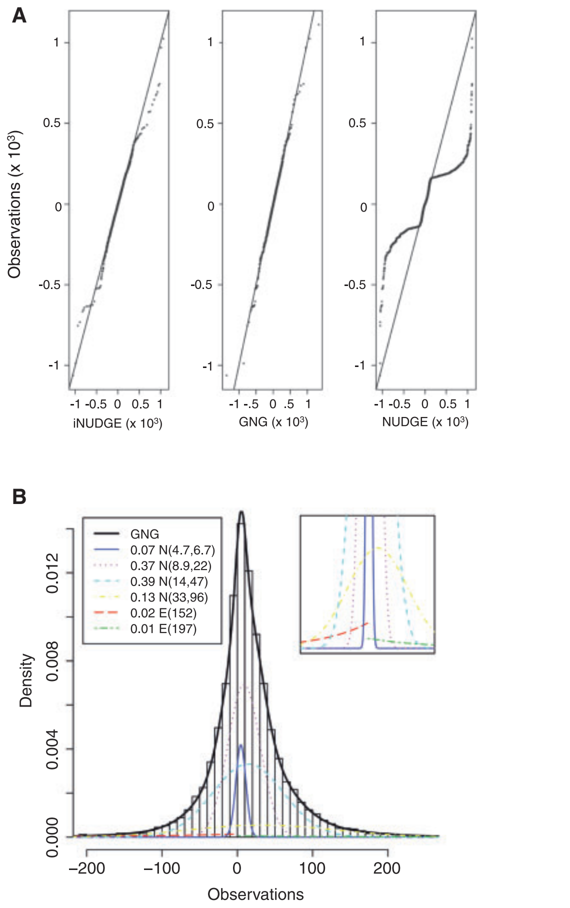

Fig. 1. (**A**) From left to right shows QQ-plot of iNUDGE, GNG and NUDGE. (**B**) Plot of the best mixture model (GNG) superimposed on the histogram of normalized data. N(.,.) stands for normal component with mean and SD. E(.) stands for exponential component with its beta parameter. Inset shows a zoomed-in plot of individual components of the model.

4 DISCUSSION

We have developed an R package (DIME) to model and make inference on ChIP-seq experiments under two different conditions. The algorithm effectively selects the model that provides the best fit to the normalized data, which lead to statistical inferences with high sensitivity and specificity. DIME can be easily combined with other R or Bioconductor packages to perform upstream and downstream analysis of ChIP-seq data. Furthermore, even though DIME is developed to fit and classify ChIP-seq data, it is highly applicable to other high-throughput data as well especially given its ensemble nature.

Funding: This work was supported in part by National Cancer Institute (grant U54CA113001) and National Science Foundation (grant DMS-1042946).

Conflict of Interest: none declared.

REFERENCES

Dean,N. and Raftery,A.E. (2005) Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics*, **6**, 173.

Hastie,T. *et al.* (2009) *The elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edn. Springer, New York.

Khalili,A. *et al.* (2009) A robust unified approach to analyzing methylation and gene expression data. *Comput. Stat. Data Anal.*, **53**, 1701–1710.

Laajala,T. *et al.* (2009) A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC Genomics*, **10**, 618.

Taslim,C. *et al.* (2009) Comparative study on chip-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**, 2334–2340.