BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 8 2012, pages 1164–1165
doi:10.1093/bioinformatics/bts087

Genome analysis

Advance Access publication February 17, 2012

GMD: measuring the distance between histograms with applications on high-throughput sequencing reads

Xiaobei Zhao* and Albin Sandelin*

Bioinformatics Centre, Department of Biology and Biotech Research and Innovation Centre, Copenhagen University, Ole Maaløes Vej 5, DK-2200, Copenhagen, Denmark

Associate Editor: Alex Bateman

ABSTRACT

Summary: GMD (generalized minimum distance of distributions) is an R package to assess the similarity between spatial distributions of read-based sequencing data such as ChIP-seq and RNA-seq. GMD calculates the optimal distance between pairs of normalized signal distributions, optionally sliding one distribution over the other to 'align' the distributions. GMD also provides graphical and downstream clustering tools.

Availability: The R package GMD source code is available at http://cran.r-project.org/web/packages/GMD/ under GPL license

Contact: xiaobei@binf.ku.dk; mailto:albin@binf.ku.dk

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

Received on November 25, 2011; revised on February 6, 2012; accepted on February 13, 2012

1 INTRODUCTION

High-throughput DNA sequencers are now a cornerstone in genomics and can be used for inferring expression, splicing and non-coding isoforms (RNA-seq), transcription start sites (TSSs) (CAGE), transcription factor binding sites, chromatin status (ChIP-seq) and more. Aside from typical tasks of read mapping/aligning to a reference or *de novo* assembling, differential expression, counting and peak calling, variant and mutation calling, comparing the distributions of DNA reads between different genomic loci or between different biological features (e.g. histone modifications) is also often meaningful. This is because the spatial distributions of such reads often indicate biological features; for example, ChIP-seq experiments targeting H3K4me3 histone modifications often aggregate in characteristic double peaks around TSSs, while the H3K36me3 mark increase from TSS to termination site (Barski *et al*., 2007). Similar examples exist for TSS (Carninci *et al*., 2006) and small RNA data (Valen *et al*., 2011).

To do this systematically, a measure of similarity between distributions is necessary. Such measures should ideally be true metrics, have as few parameters as possible, be computationally efficient and also make biological sense to end-users. Here, we present such a measure, generalized minimum distance of distributions (GMD), based on MDPA (minimum difference of pair assignment, Cha and Srihari, 2002). It can compare two empirical distributions of categorical data, which we refer as histograms (Supplementary Material). Considering two normalized histograms A and B, GMD measures their similarity by counting the necessary 'shifts' of elements between the bins that have to be performed to transform distribution A into distribution B. Both GMD and MDPA have been implemented in C to interface with R for computational efficiency.

*To whom correspondence should be addressed.

2 ALGORITHM

The heart of GMD is the comparison of two histograms, typically originating from counting DNA reads from biological experiments (e.g. ChIP) mapped to two limited genomic regions, which have been normalized so that they sum to 1. The measuring of GMD of two normalized ordinal type histograms A and B is formulated into a process we call *gmd*, with an option *sliding*, indicating whether the optimal alignment should be searched for partial alignment.

(**a**)

```
PROCEDURE mdpa(O,T)
  E=O-T
  S=CUMULATIVE_SUM(E)
  RETURN SUM(S)
END PROCEDURE
```

(**b**)

```
PROCEDURE gmd(A,B)
  b1=LENGTH(A)
  b2=LENGTH(B)
  LET Q=CONCATENATE(Z(b1),B,Z(b1))
  LET D=()
  FOR i FROM 1 TO (b1+b2+1) BY 1
    P=CONCATENATE(Z(i-1),A,Z(b1+b2-i+1))
    LET d=mdpa(P,Q)
    LET D=CONCATENATE(D,d)
  ENDFOR
  RETURN MIN(D)
END PROCEDURE
```

Fig. 1. The process *gmd* and its building block *mdpa*. The *gmd* process complements the *mdpa* by sliding and padding relevant distributions with zero counts and finding the optimal alignment, where $Z(n)$ denotes a zero vector of length n. (**a**) the process *mdpa* for MDPA, where O and T denote two ordinal type histograms, with equal size of bins and equal overall mass. (**b**) the process *gmd* (with the option *sliding* on) for GMD, where A and B denote two normalized ordinal type histograms, with bin size b_1 and b_2, respectively, ($b_1 \leq b_2$). This algorithm runs in $\mathcal{O}(b)$ or $\mathcal{O}(b^2)$ with *sliding* off or on respectively, where b is the size of the distributions. (see Supplementary Material for details)

3 RESULTS

Here, we show two examples of comparing distributions of CAGE and ChIP-seq data around TSSs, and an example of multiple pairwise comparison and the cluster utilities in GMD.

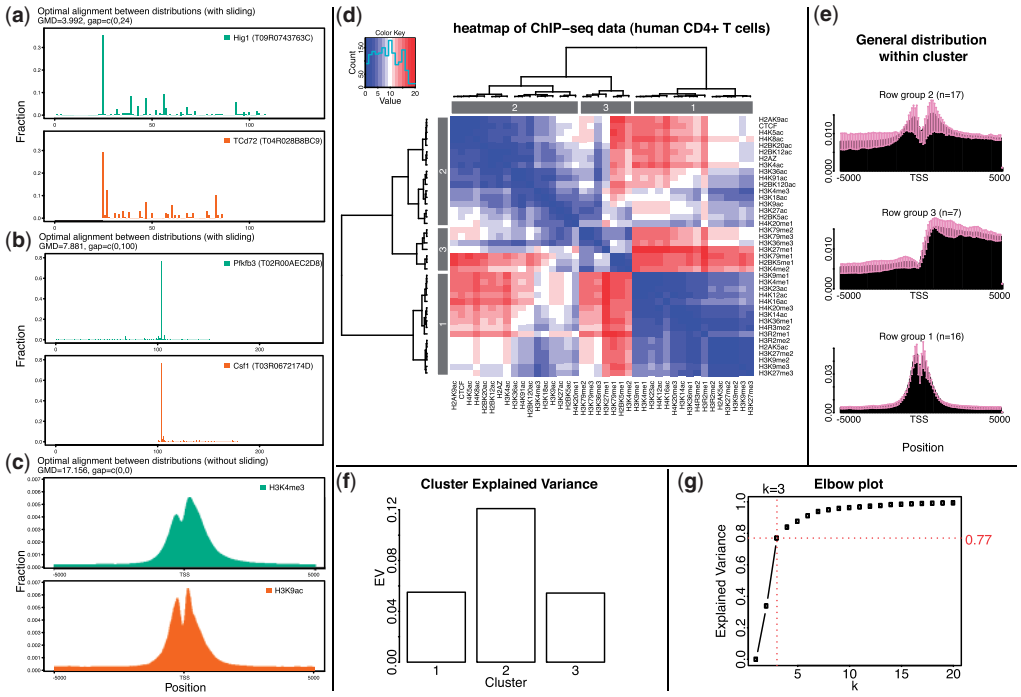

Fig. 2. Example analysis and output from GMD. (**a**–**b**) CAGE TSS usage distributions from different parts of the mouse genome are aligned and compared: (**a**) shows the comparison of two peaked distributions while (**b**) shows the alignment of two broader distributions. *X*-axis shows nucleotide positions and the *Y*-axis shows the fraction of CAGE reads. The distance and alignment shifts are shown on top of each image. Gene names corresponding to each promoter are shown. (**c**) Comparison of the mean footprint of two ChIP datasets around all TSSs. (**d**) Extension of the analysis in (c) to pairwise comparisons of 40 ChIP distributions, represented as a clustered heatmap where each row and column is one ChIP footprint around the TSS. The blue color represents high similarity over TSS regions. Three larger clusters are proposed, shown as numbered blocks on the sides. (**e**) Summary histograms of each cluster, with pink confidence intervals. *X*-axis denote the ±5000 nt region around TSSs. (**f**) Clustering statistics showing the within-cluster sum-of-squares (WSSs). Smaller WSSs indicate higher cluster cohesion. (**g**) Explained variance of the clustering model as a function of the number of clusters k [(an 'elbow' plot (Thorndike, 1953))]. All of these plots are taken directly from the `GMD` output.

Case study I: Transcription initiation distributions (CAGE): CAGE captures and maps mRNA 5' ends to the genome and can be used to infer TSS usage (Carninci *et al*., 2006). In Figure 2a–b, we measure the similarity between two TSS distributions, corresponding to two core promoters, taken from (Zhao *et al*., 2011).

Case study II: Histone modification distributions (ChIP-seq): Histograms that are the result of averaging over a whole dataset can also be compared. In Figure 1c, we have compared the mean footprint over all TSSs of histone modifications H3K4me3 and H3K9ac from ChIP data (Barski *et al*., 2007; Mikkelsen *et al*., 2007). Figure 1d–f extends this to pairwise comparisons between 40 ChIP sets, along with summary statistics of the resulting clusters or the overall clustering quality.

4 CONCLUSION

The GMD metric differs from common distance measurements (e.g. Euclidean norm, KL divergence) since the bins in the histogram will not be treated as independent and the similarity between non-overlapping bins is taken into account (discussed in Cha and Srihari, 2002) and partial alignment is allowed. This makes sense in many applications when we have little knowledge about bin-to-bin correspondence between distributions. The `GMD` package together with its vignette and manual makes systematic studies comparing distributions possible with limited R knowledge.

Funding: This work was supported by the Novo Nordisk Foundation, and European Research Council (FP7/2007-2013; 204135) and by the Lundbeck Foundation.

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al*. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Carninci,P. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

Cha,S. and Srihari,S. (2002) On measuring the distance between histograms. *Pattern Recogn.*, **35**, 1355–1370.

Mikkelsen,T.S. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Thorndike,R.L. (1953) Who belong in the Family? *Psychometrika*, **18**, 267–276

Valen,E. *et al*. (2011) Biogenic mechanisms and utilization of small rnas derived from human protein-coding genes. *Nat. Struct. Mol. Biol.*, **18**, 1075–1082.

Zhao,X. *et al*. (2011) Systematic clustering of transcription start site landscapes. *PLoS ONE*, **6**, e23409.