BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 17 2011, pages 2446–2447
doi:10.1093/bioinformatics/btr404

Gene expression
Advance Access publication July 7, 2011

PREDA: an R-package to identify regional variations in genomic data

Francesco Ferrari[1], Aldo Solari[2], Cristina Battaglia[3] and Silvio Bicciato[1,*]

[1]Department of Biomedical Sciences, Center for Genome Research, University of Modena and Reggio Emilia, Modena, [2]Department of Statistics, University of Milano-Bicocca and [3]Department of Biomedical Sciences and Technologies, University of Milano, Milano, Italy

Associate Editor: John Quackenbush

ABSTRACT

Summary: Chromosomal patterns of genomic signals represent molecular fingerprints that may reveal how the local structural organization of a genome impacts the functional control mechanisms. Thus, the integrative analysis of multiple sources of genomic data and information deepens the resolution and enhances the interpretation of stand-alone high-throughput data. In this note, we present *PREDA* (Position RElated Data Analysis), an R package for detecting regional variations in genomics data. *PREDA* identifies relevant chromosomal patterns in high-throughput data using a smoothing approach that accounts for distance and density variability of genomics features. Custom-designed data structures allow efficiently managing diverse signals in different genomes. A variety of smoothing functions and statistics empower flexible and robust workflows. The modularity of package design allows an easy deployment of custom analytical pipelines. Tabular and graphical representations facilitate downstream biological interpretation of results.

Availability: *PREDA* is available in Bioconductor and at http://www.xlab.unimo.it/PREDA.

Contact: silvio.bicciato@unimore.it

Supplementary information: Supplementary information is available at *Bioinformatics* online.

Received on April 6, 2011; revised on June 6, 2011; accepted on June 23, 2011

1 INTRODUCTION

High-throughput technologies allow the complete exploration of DNA and RNA molecules at genome-wide scale, thus offering the opportunity to decipher the control mechanisms underlying the functional utilization of a genome. Although hundred of studies fully demonstrated the relevance of bioinformatics in analyzing each data type alone, fewer approaches have been proposed for the integrative analysis of genomics signals and genome structural organization. Some of these methods specifically address the identification of gene expression patterns at the chromosomal level (Callegaro *et al*., 2006; Pollack *et al*., 2002; Toedling *et al*., 2005), others the integrative analysis of paired DNA copy number and gene expression data (Bicciato *et al*., 2009; Lahti *et al*., 2009; Salari *et al*., 2010; Schafer *et al*., 2009).

Here, we describe *PREDA*, an R/Bioconductor package for detecting regional variations of genomic features from the integrative analysis of high-throughput data and genome local structural organization. The program implements a generalized version of the integrative strategy first introduced by Toedling *et al*. (2005) and further improved by Callegaro *et al*. (2006) to identify and prioritize targets for functional studies from diverse type of high-throughput data in different genomes (Bicciato *et al*., 2009; Coppe *et al*., 2009; Ferrari *et al*., 2007; Nie *et al*., 2009; Peano *et al*., 2007). *PREDA* integrates high-throughput signals and structural information using a non-linear kernel regression with adaptive bandwidth that efficiently accounts for the density variability of genomics data. The integrative analysis is performed through a modular and flexible framework that, accommodating different types of smoothing functions and statistics, facilitates its adoption in a broad spectrum of genomics studies. The potential applications range from the study of pathological processes involving genome structure modifications, as for chromosomal amplifications and deletions in cancer (e.g. integrative analysis of DNA copy number and gene expression), to the analysis of physiological mechanisms regulating genome utilization (e.g. identification of genes co-localized and co-expressed; analysis of tiling arrays). A detailed description and discussion of the package is reported in the Supplementary Material.

2 IMPLEMENTATION

PREDA has been implemented as a package for R statistical environment and is compliant with Bioconductor standards for bioinformatics packages development and documentation.

Data structures: the package takes advantage of basic Bioconductor objects classes to facilitate the management of input data and to allow an easier integration of the *PREDA* pipeline in other workflows. Moreover, additional custom S4 data structures have been defined for a more efficient management of data and genomic information and to facilitate further customization of the analysis workflow.

Modular framework: the basic computational framework has been developed focusing on modularity. Therefore, it can be easily extended if custom analytical pipelines are required for more complex or specialized purposes, or to integrate *PREDA* analysis into other computational pipelines.

Parallel computing: since the analytical procedure is time consuming due to repeated cycles of data permutations and

[*]To whom correspondence should be addressed.

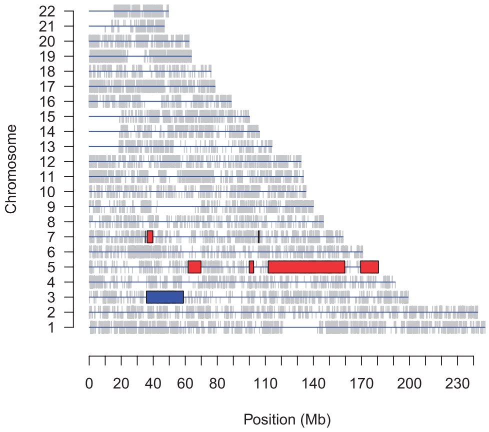

Fig. 1. Amplified (red) and deleted (blue) regions identified by *PREDA* through the integrative analysis of DNA copy number and gene expression data in clear cell renal carcinomas (as described in the package tutorial).

smoothing, a parallelized version of the algorithm has been implemented to reduce the computation time. The parallelization is based on Rmpi (R interface to widespread MPI parallel libraries), which allows taking advantage both of high performance computing systems and of modern multicore processors that are currently available in common desktop computers. The computing time scales almost linearly up to elevate numbers of CPUs (Supplementary Material).

Documentation: the *PREDA* specific S4 classes are discussed and detailed in package vignettes and in a step-by-step tutorial for R users with different levels of experience. In addition, each function of the package has a specific documentation page, according to Bioconductor standards.

3 METHODS

The core of the *PREDA* procedure consists of three major steps:

(1) computation of the observed statistic on each position (e.g. genes, transcripts, any other genomic feature);

(2) non-linear regression smoothing of observed statistics along the genomic coordinate;

(3) construction and smoothing of the expected statistics. The expected statistic is obtained permuting the position-related statistics and allows empirically estimating the local significance of peaks in the observed smoothed statistics (empirical *P*-values).

The empirical *P*-values resulting from permutations analysis are subsequently adjusted to control False Discovery Rate (FDR) and used to identify significant genomic regions (Fig. 1). Several functions have been implemented to facilitate the management of different input data (e.g. genomics signals and annotations) and to enhance the extraction of biologically relevant information from the output results (e.g. significant genomic regions). In addition, several alternative methods for data smoothing and quantification of statistics have been implemented to allow full flexibility and applicability to data obtained from different technologies (Supplementary Materaial). *PREDA* supports a variety of tabular and graphical options for visualizing results at the level of single chromosomes or of the entire genome (Fig. 1). The package contains some predefined workflows (implemented as wrapper functions) to perform the multiple steps of *PREDA* in a facilitated way.

4 RESULTS

PREDA generalizes and strengthens previously proposed methods for the integrative analysis of high-throughput data and genome structural information (Bicciato *et al.*, 2009; Callegaro *et al.*, 2006; Toedling *et al.*, 2005). The modularity of the package allows implementing different type of smoothing functions and statistics, thus facilitating its application to different genomics data and genomes. The core components of the method, i.e. (i) the data smoothing with non-linear regression functions; (ii) the permutation schema to evaluate the significance of results; (iii) the statistics computed on each gene and used as input for the analysis, have been rendered of general applicability and enhanced over prior implementations. The modifications aimed at improving the sensitivity of the procedure in detecting small significant genomic regions and at controlling false positive results. The methodological refinements were tested on simulated datasets and the results are discussed in details in the Supplementary Material. Results indicate that the overall procedure for position-related data analysis is robust to modifications of input statistics and can be adopted in a number of different biological problems. However, simulations revealed that selecting the optimal smoothing strategy might be critical to improve sensitivity and reduce the FDR. As such, several alternative options for data smoothing have been implemented in the *PREDA* package along with a flexible and modular framework of R functions that would allow end users to define custom smoothing methods, when needed.

ACKNOWLEDGEMENTS

The authors acknowledge the scientific support of Judith Boer, Leiden University Medical Center, The Netherlands.

Funding: HPC-EUROPA2 Project (211437); MIUR (PRIN 2007Y84HTJ, FIRB RBNE01HCKF and RBRN07BMCT); AIRC Special Program Molecular Clinical Oncology 5 per Mille 2010.

Conflict of Interest: none declared.

REFERENCES

Bicciato,S. *et al*. (2009) A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Res.*, **37**, 5057–5070.

Callegaro,A. *et al*. (2006) A locally adaptive statistical procedure (lap) to identify differentially expressed chromosomal regions. *Bioinformatics*, **22**, 2658–2666.

Coppe,A. *et al*. (2009) Motif discovery in promoters of genes co-localized and co-expressed during myeloid cells differentiation. *Nucleic Acids Res.*, **37**, 533–549.

Ferrari,F. *et al*. (2007) Genomic expression during human myelopoiesis. *BMC Genomics*, **8**, 264.

Lahti,L. *et al*. (2009) Dependency detection with similarity constraints. In Adali,T. *et al.* (eds) *Proceedings of the 2009 IEEE International Workshop on Machine Learning for Signal Processing XIX*, IEEE, pp. 89–94.

Nie,H. *et al*. (2009) Microarray data mining using Bioconductor packages. *BMC Proc.*, **3** (Suppl. 4), S9.

Peano,C. *et al*. (2007) Complete gene expression profiling of Saccharopolyspora erythraea using genechip DNA microarrays. *Microb. Cell Fact.*, **6**, 37.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Salari,K. *et al.* (2010) DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, **26**, 414–416.

Schäfer,M. *et al.* (2009) Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, **25**, 3228–3235.

Toedling,J. *et al*. (2005) Macat-microarray chromosome analysis tool. *Bioinformatics*, **21**, 2112–2113.