BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 12 2012, pages 1643–1644
doi:10.1093/bioinformatics/bts201

Genome analysis

Advance Access publication April 23, 2012

NuST: analysis of the interplay between nucleoid organization and gene expression

Vittore F. Scolari[1,2,3,*], Mina Zarei[1,2,4], Matteo Osella[1,2] and Marco Cosentino Lagomarsino[1,2,5]

[1]Genomic Physics Group, UMR 7238 CNRS Génomique des Microorganismes, [2]Université Pierre et Marie Curie, 15 rue de L'École de Médecine, 75006, Paris, France, [3]NCBS, Bangalore, India, [4]Dip. Fisica, Università di Milano, Milano, Italy and [5]Dip. Fisica, Università di Torino, Torino, Italy

Associate Editor: Martin Bishop

ABSTRACT

Summary: Different experimental results suggest the presence of an interplay between global transcriptional regulation and chromosome spatial organization in bacteria. The identification and clear visualization of spatial clusters of contiguous genes targeted by specific DNA-binding proteins or sensitive to nucleoid perturbations can elucidate links between nucleoid structure and gene expression patterns. Similarly, statistical analysis to assess correlations between results from independent experiments can provide the integrated analysis needed in this line of research. NuST (Nucleoid Survey tools), based on the *Escherichia coli* genome, gives the non-expert the possibility to analyze the aggregation of genes or loci sets along the genome coordinate, at different scales of observation. It is useful to discover correlations between different sources of data (e.g. expression, binding or genomic data) and genome organization. A user can use it on datasets in the form of gene lists coming from his/her own experiments or bioinformatic analyses, but also make use of the internal database, which collects data from many published studies.

Availability and Implementation: NuST is a web server (available at http://www.lgm.upmc.fr/nust/). The website is implemented in PHP, SQLite and Ajax, with all major browsers supported, while the core algorithms are optimized and implemented in C. NuST has an extensive help page and provides a direct visualization of results as well as different downloadable file formats. A template Perl code for automated access to the web server can be downloaded at http://www.lgm.upmc.fr/nust/downloads/, in order to allow the users to use NuST in systematic bioinformatic analyses.

Contact: vittore.scolari@upmc.fr

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

Received on February 22, 2012; revised on April 12, 2012; accepted on April 16, 2012

The orchestration of coordinated global changes in the transcriptional program is at the basis of bacterial adaptation to environments and stresses. These changes depend on a large regulatory network mediated by specific binding of transcription factors, as well as on the physical organization of the chromosome, which affects the expression of large gene sets (Rimsky and Travers, 2011). The bacterial DNA is condensed in a compact DNA–protein complex called 'nucleoid', whose transcriptional activity depends on nucleoid-related factors including the degree of supercoiling (Travers and Muskhelishvili, 2005) and the specific and non-specific binding of nucleoid associated proteins (NAPs) such as Fis, H-NS and HU (Dame *et al.*, 2011). While the total level of supercoiling is controlled by enzymes like gyrases and topoisomerases (Travers and Muskhelishvili, 2005), NAPs like H-NS and Fis are believed to stabilize locally DNA loops (Luijsterburg *et al.*, 2006). The nucleoid structure varies at different scales, from DNA supercoiled loops ∼10 kb long (Postow *et al.*, 2004) to large compartments organizing the genome in four 'macrodomains' (Dame *et al.*, 2011; Valens *et al.*, 2004). The effects of transcriptional network and nucleoid on large-scale gene expression are coupled. For example, many NAPs are also specific transcription factors, and affect the expression of targets both directly and through the conformational changes that they can induce on the chromosome.

The complex interplay between chromosome organization and gene expression requires integration of data from different high- and low-throughput experiments with statistical analysis at multiple scales. The web server described here is an effort to fill this gap. Since many of the nucleoid structural features (e.g. supercoil domains and macrodomains) involve contiguous genomic regions, a typical hallmark of nucleoid-mediated regulation is aggregation along the chromosome of genes having specific properties (Mathelier and Carbone, 2010; Scolari *et al.*, 2011; Sobetzko *et al.*, 2012). The main tool is able to identify significant linear aggregation clusters of a gene set, considering different observation scales (Scolari *et al.*, 2011). The web server is currently based on *Escherichia coli*. An extensive step-by-step documentation introduces analyses that can be performed using the web server and is divided into an introductory help page, and a 'learn by example' page guiding the user through the interpretation of the results and the choice of the parameters.

The input datasets are single column text files with one gene ID for each row. Standard gene IDs are Regulon DB database (Gama-Castro *et al.*, 2011) (for different gene IDs the server proposes synonyms). Sample datasets as well as the complete list of accepted synonyms can be obtained from the 'Download' page. Loaded data sets are stored in the 'Personal' part of the internal database and can be accessed for further analysis. They are deleted at the end of each anonymous session. A login (obtained sending an email to the administrator) allows to keep personal data for multiple sessions. The 'Common' part of the database contains datasets from the literature, organized by type of data and experimental technique.

*To whom correspondence should be addressed.

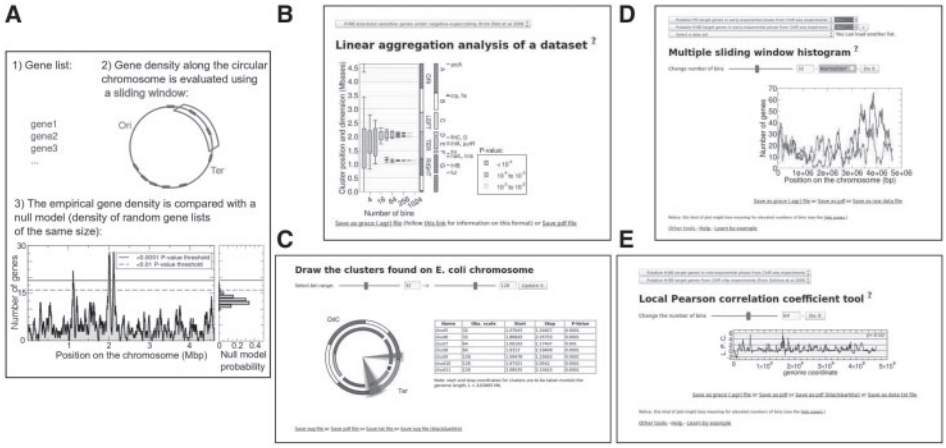

Fig. 1. Main features of NuST. (**A**) One-dimensional aggregation of gene set. Starting from a list of genes, their density is evaluated using a sliding window of different size and compared to a shuffling null model to extract the significant clusters. (**B**) Diagram of the statistically significant linear aggregation clusters. The box indicates the position of the peak at a given scale of analysis (x axis) while the whisker indicates the maximal extension of the cluster, with color-coded P-values. The right panel reports the positions of chromosomal macrodomains (Dame *et al.*, 2011; Valens *et al.*, 2004), chromosomal sectors defined in (Mathelier and Carbone, 2010), and the names of some important genes. (**C**) Circular representation of the most significant clusters found for the same data set of Fig. 1B. The outer colored ring represents macrodomains (Dame *et al.*, 2011; Valens *et al.*, 2004) while the inner ring represents chromosomal sectors defined by Mathelier and Carbone (Mathelier and Carbone, 2010). The table reports the cluster ID, the scale of observation, the cluster coordinates in megabases and the P-value associated to each cluster. (**D**) Overlaid histogram of two datasets in the common database. The input files belonging to the common datasets can be read in the upper part of the plot. The normalized number of genes located inside a window centered in each chromosomal position is shown. In this example the window size is fixed to $L/16$, where L is the total length of the genome. (**E**) Local Pearson correlation function between sliding windows histograms. The (customizable) window size is $L/32$, where L is the total length of the genome. The plot legend reports the total Pearson correlation coefficient. In this case, although globally there is no correlation, specific regions can present cooccurrence or mutual exclusion.

The main features of the server are described in Fig. 1. The 'linear aggregation' analysis (Fig. 1A) detects significant aggregation along the genome coordinate. The output is directly visualized on the website with two bitmap pictures and a table (Fig. 1B). An alternative graphical representation shows the statistically significant clusters as colored wedges (Fig. 1C). Sliding-window histograms (Fig. 1D) allow the comparison of the gene density of different datasets. Finally, a tool evaluates the local contribution to the Pearson correlation coefficient along the chromosome between the gene densities of two gene sets (Fig. 1E). Since averages and standard deviations are calculated along the whole genome, the global Pearson correlation coefficient is a number between -1 (linear anticorrelation) and $+1$ (linear correlation). The local product does not have this constraint, but represents a measure of positive or negative correlation (affected by the deviation of the two densities from their global averages). Note that the global Pearson correlation coefficient can change with observation scale, as the sliding-window histograms may be different.

To the best of our knowledge the NuST web server is the first freely available computational tool performing multi-scale analysis of the linear aggregation of specific gene sets along the chromosome. It is designed for the investigation of links between bacterial chromosome organization and the global gene expression program, a problem in which such multi-scale approach and comparison/integration of different kinds of data can be highly informative.

ACKNOWLEDGEMENTS

We would like to thank B. Sclavi, P. Cicuta, M. Cereda, M. Babu and S. Wielgoss for useful feedback.

Funding: This work was supported by the International Human Frontier Science Program Organization, grant [RGY0069/2009-C].

Conflict of Interest: none declared.

REFERENCES

Dame,R.T., Kalmykowa,O.J., and Grainger,D.C. (2011) Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. *PLoS Genet*., **7**, e1002123.

Gama-Castro,S. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*., **39**, D98–D105.

Luijsterburg,M.S. *et al.* (2006) The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: a molecular perspective. *J. Struct. Biol*., **156**, 262–272.

Mathelier,A. and Carbone,A. (2010) Chromosomal periodicity and positional networks of genes in Escherichia coli. *Mol. Syst. Biol*., **6**, 366.

Postow,L. *et al.* (2004) Topological domain structure of the Escherichia coli chromosome. *Genes. Dev*., **18**, 1766–1779.

Rimsky,S. and Travers,A. (2011) Pervasive regulation of nucleoid structure and function by nucleoid-associated proteins. *Curr. Opin. Microbiol*., **14**, 136–141.

Scolari,V.F. *et al.* (2011) Gene clusters reflecting macrodomain structure respond to nucleoid perturbations. *Mol. Biosyst*., **7**, 878–888.

Sobetzko,P. *et al.* (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl. Acad. Sci. USA*., **109**, E42–E50.

Travers,A. and Muskhelishvili,G. (2005) DNA supercoiling – a global transcriptional regulator for enterobacterial growth? *Nat. Rev. Microbiol*., **3**, 157–169.

Valens, M. *et al.* (2004) Macrodomain organization of the Escherichia coli chromosome. *EMBO J*., **23**, 4330–4341.