BIOINFORMATICS

Vol. 00 no. 00
Pages 1–2

Constitution Analysis

Israel T. Silva [1,3,*], and Rafael A. Rosales [2]

[1]Laboratory of Molecular Immunology, The Rockefeller University, 1230 York Avenue, New York, NY 10065
[2]Departamento de Computação e Matemática, Universidade de São Paulo. Av. Bandeirantes, 3900, Ribeirão Preto, CEP 14049-901, SP, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:

Results: https://github.com/someone/rich.

Contact: someone@somewhere.world

1 INTRODUCTION

The genome is targeted by a sophisticated and highly coordinated series of molecular events. Among these events, aberrant DNA methylation patterns in human malignancy De *et al.* (2013), somatic retrotransposition in human cancers Lee *et al.* (2012), AID-dependent chromosomal translocations (Klein, 2011) and HIV integration (Cohn, 2014), which arrives throughout DNA, are not randomly distributed but instead associated with chromosomal regions and contributes to disrupt the integrity of the genome and human disease.

As result, these regions represents a genomic context in which are associate with multiple underlying mechanisms. The motif-based sequence analysis is the starting point to aim potential binding site of cis-regulatory elements associated. Nevertheless, the inherent low signal/noise ratio in sequence-based motif discovery is a limitation to detect a nucleic acid sequence pattern that has some biological significance. Moreover, these events may recognise a structural feature, rather than a specific sequence motif.

Others approaches were introduced to detect functional regions using methods for computing sequence complexities Jin *et al.* (2014); Koslicki (2011). In these methods, the complexity is measured by the entropy which evaluates the randomness of DNA sequence. In particular, topological entropy has been applied to compute the complexity of introns, exons and promoter regions. Due to the finite sample and high-dimensionality problems, efforts aimed to overcome these problems are put forward Koslicki (2011).

However, how exactly the pattern nucleotide composition could influence the selection of target site selection are not well understood. To further characterize at a genome-wide scale these regions, we introduce a new method to provide a quantitative measure of the differential spectra of k-mers (DNA 'words' of length k) inside target DNA.

*to whom correspondence should be addressed

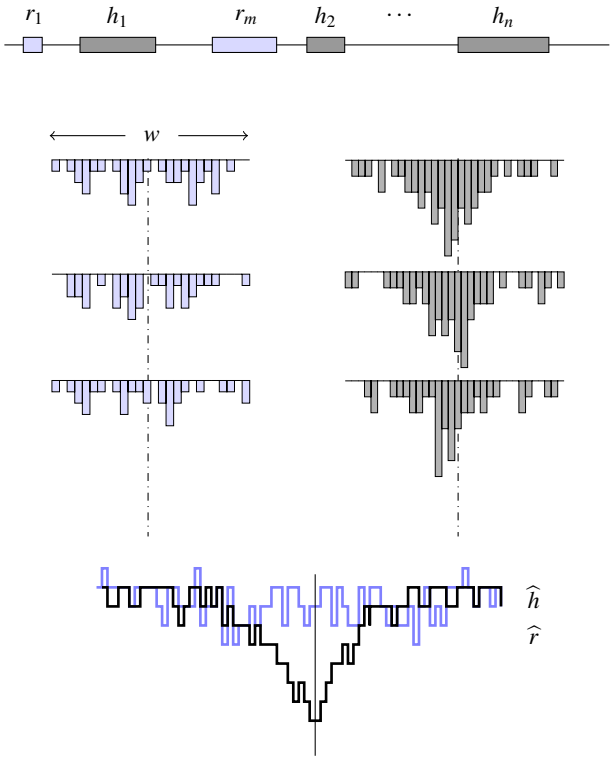

Fig. 1. The method

2 METHODS

We specialise to regions conforming to viral insertion hotspots, but the method is clearly not restricted to this application. Insertion hotspots are detected by using scan statistics as described in Silva *et al.* (2014). Let $\{h_1, \ldots, h_n\}$, bet a set of insertion hotspots and let w be the length (in base pairs) of the longest such segments. The coordinates of all the hotspots were extended at bot extremes such to match exactly this length. The resulting set of segments were partitioned into k evenly spaced bins of length $\rho = w/k$. Denote by h_{ij}, $1 \leq j \leq k$, the jth interval of the ith segment. Let $\{r_1, \ldots, r_n\}$, be a set segments of width w, uniformly distributed along the genome but with no intersection with h_i, $1 \leq i \leq n$. Also, denote by r_{ij} the jth interval of the ith segment.

Let $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ and $\mathcal{S} \in \mathcal{A}^{\ell}$, be a string of length $\ell \leq \rho$. Let ξ_{ij} and η_{ij} for $1 \leq i \leq n$, $1 \leq j \leq k$, be random variables defined as

$$\xi_{ij} = \begin{cases} 1, & \text{if } \mathcal{S} \in h_{ij} \\ 0, & \text{otherwhise} \end{cases} \quad \text{and} \quad \eta_{ij} = \begin{cases} 1, & \text{if } \mathcal{S} \in r_{ij} \\ 0, & \text{otherwhise} \end{cases}.$$

Consider the counts

$$X_j = \sum_{i=1}^{n} \xi_{ij}, \qquad Y_j = \sum_{i=1}^{n} \eta_{ij}$$

that is X_j and Y_j reccord the number of times that $\mathcal{S}$ occurs along each bin of a hotspot and a randomly chosen region respectively. In order to asses if the distribution of $\mathcal{S}$ varies differently across the bins of a hotspot region, we may consider the following test. Let

$$X^2 = \sum_{j=1}^{k} (X_j - Y_j)^2 / Y_j.$$

Provided the number of counts in each of the k bins is sufficiently large, X^2 is a χ^2 random variable with $k-1$ degrees of freedom. This provides then the sampling distribution to perform Pearson's χ^2 test for a $2 \times k$ contingency table, with H_0 and H_a ...

REFERENCES

De, S., Shaknovich, R., Riester, M., Elemento, O., Geng, H., Kormaksson, M., Jiang, Y., Woolcock, B., Johnson, N., Polo, J. M., Cerchietti, L., Gascoyne, R. D., Melnick, A., and Michor, F. (2013). Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity. *PLoS Genet.*, **9**(1), e1003137.

Jin, S., Tan, R., Jiang, Q., Xu, L., Peng, J., and Wang, Y. (2014). A generalized topological entropy for analyzing the complexity of DNA sequences. *PLoS ONE*, **9**(2), e88519.

Koslicki, D. (2011). Topological entropy of DNA sequences. *Bioinformatics*, **27**(8), 1061–1067.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R., Lee, C., Kharchenko, P. V., and Park, P. J. (2012). Landscape of somatic retrotransposition in human cancers. *Science*, **337**(6097), 967–971.

Silva, I. T., Rosales, R. A., Holanda, A. J., Nussenzweig, M., and Jankovic, M. (2014). Identification of chromosomal translocation hotspots via scan statistics. *Bioinformatics*, **2**, 1–8.