Fast assessment of the correlation between different coverage-like genomic features and of its statistical significance

Elena Stavrovskaya*, Alexander Favorov, and Andrey Mironov

Faculty of Bioengineering and Bioinformatics,
Lomonosov Moscow State University,
Leninskie gory 1-73, Moscow, 119992, Russia
Institute for Information Transmission Problems
Bolshoy Karetny per. 19, Moscow, 127994, Russia
Vavilov Institute of Genral Genetics RAS,
Gubkina str. 3, Moscow, 119333, Russia
State Scientific Center Genetika,
1-st Dorozhniy pr., 1, Moscow, 117545, Russia
Johns Hopkins University School of Medicine,
550 N Broadway ste 1103 Baltimore, MD 21205 USA

Abstract. The modern high-throughput sequencing methods provide massive amounts of genome-focused, DNA-positioned data. This data is often represented as a function of the DNA coordinate (e.g. coverage). The genome- or chromosome-wide correlations between data from different sources may provide information about functional biological interrelation of the investigated features, e.g., trancription and histone modification. The task to compute the correlation was already successfully solved for interval annotations ([1]) as well as for coverage (functional) data ([2], [3], [4]). The key idea of the correlation studies is that two features that are similarly distributed along a chromosome may be functionally related. The point we are addressing here is a that peaks of dependent functional features can be located in a similar, although somewhat different, way. To account for these similarities, we propose here a fast method for calculation of kerneled correlation between two numeric annotations of the genome. The kernel represents the mutual position of related features; e.g., a Gaussian shape corresponds to 'somewhere around', etc.

Keywords: positional correlation, coverage NGS data, FFT

Introduction

Current experimental techniques generate large amounts of data related to the genome. This information is generally aggregated in publicly available storages, such as the UCSC Genome Browser [5], Epigenomic Roadmap [6], or ENCODE

* corresponding author

[7]. The information that is linked to the genome (genomic feature) usually is represented in one of two ways: either as intervals (for example, the location of genes, SINE elements, CpG-islands, etc.) or as a continuous distribution (for example, the number of reads in RNA-seq, or the conservation score of the sequence).

The relation between different kinds of such data can facilitate understanding of the underlying biological processes. Comparative genomic and epigenetic studies that are based on this idea (see, e.g., [8–11]) and specialized tools for data integration and visual analysis [12–15] rapidly emerge nowadays.

A common method for comparison of genomic features is based on their interval representation [4]. In this approach, the association is measured on the basis of weight overlap of intervals [4, 1] or of inter-interval distances [16, 17, 1]. The latter approach can reveal a distant interaction that occurs between genomic features. All the interval approaches binarize non-interval data before analysis; therefore, the results are dependent on parameter choice.

Several statistical methods that estimate the association between genome-wide numerical features have been proposed recently. They are based on different measures of similarity, e.g correlation [18], the number of clusters of transcription factor binding sites [19], or distances [20, 21]. The likelihood that the observed values were obtained by chance can be obtained either from Monte Carlo simulations [20, 21, 18] or from analytical assessment using parametric models [22]. A complete review is presented in [23].

Here, we propose a universal method for comparison of genomic features. The compared features can be either discrete (e.g., genes or repeats) or continuous (e.g., the level of histone methylation CHiP-Seq signal). Moreover, the method accounts both for local and distant similarities. The method relies on an integral measure of the similarity that is calculated by the Fourier transform. The measure is calculated in a set of windows. It allows not only to make assertions about the similarity of the features across the genome, but also to detect genomic fragments with unusual behavior. The method is tested using a comparison of histone modification marks and the mRNA-seq signal.

Methods

A generalisation of Jaccard measure

Assume two predefined sets of intervals F, G. A Jaccard measure can be used for comparison of these two sets:

$$J = \frac{F \bigcap G}{F \bigcup G}$$

Each set of intervals can be described with an indicator function:

$$f(x) = \begin{cases} 1, & x \in F \\ 0, & x \notin F \end{cases}$$

Using indicator functions, the numerator can be written as $\int f(x)g(x)dx$. The Jaccard measure detects only strong overlap of intervals. If intervals F and G are spatially close, but do not overlap, the Jaccard measure will be equal to zero. To circumvent this, we can modify numerator in the above definition as follows:

$$Q(f,g) = \int\int f(x)g(y)\rho(y-x)dxdy \tag{1}$$

where ρ is some positive kernel. In our analysis, we will ignore the denominator; we will introduce other normalization later. Equation (1) has two important features. First, in addition to the overlap, it reflects teh proximity of interval sets. Second, it can be applied to an arbitrary genome profile, rather than only to interval sets.

The kernel function $\rho(y-x)$ can have different forms. If the aim is simply to calculate the correlation between two features, the Gauss kernel can be used $\rho(z) = exp\left(-z^2/\sigma^2\right)$; parameter σ reflects the width of the kernel. A non-symmetrical kernel can be also used, for example, if we are interested in analysing the chromatin properties only in upstream regions of genes.

Fourier analysis

Let $f(x)$ and $g(x)$ be genome-based sets of features (observations). A similarity measure of these features that is based on kernel ρ is defined by Equation (1). All integrands in this equation can be presented as a Fourier series with respect to some orthonormal (complex) basis $\{\phi_i\}$:

$$f(x) = \sum_i f_i\phi_i(x); \quad g(x) = \sum_i g_i\phi_i^*(x); \quad \rho(y-x) = \sum_{i,j}\rho_{i,j}\phi_i(x)\phi_k^*(y)$$

Here, $\phi_i^*(x)$ means complex conjugation. Then the integral can be rewritten as:

$$Q(f,g) = Re\left(\int_0^L\int_0^L \sum_{i,j} f_i\phi_i(x)\phi_j^*(x)\sum_{k,l}\rho_{j,l}g_k\phi_k(t)\phi_l^*(t)dtdx\right) = Re\left(\sum_{i,k} f_i g_k \rho_{i,k}\right)$$

Let the basis function $\phi_k(x)$ to be a standard basis of Fourier transformation:

$$\phi_k(x) = e^{ikx\cdot 2\pi/L}$$

Fourier transformation of $\rho(y-x)$ is:

$$\rho(y-x) = \sum_k \rho_k \cdot e^{ik(y-x)\cdot 2\pi/L} = \sum_k \rho_k \cdot \phi_k(y)\phi_k^*(x)$$

Our similarity measure becomes:

$$\begin{aligned} Q(f,g) &= Re\left(\int\limits_0^L \int\limits_0^L f(x)g(y)\rho(y-x)dxdt\right) \\ &= Re\left(\sum_{k,l,m} f_k g_l \rho_m \int\limits_0^L \phi_k(x)\phi_m^*(x) \int\limits_0^L \phi_l^*(y)\phi_m(y)dydx\right) \end{aligned}$$

Here, we use the property of basis functions $\phi_k(y-x) = \phi(y)\cdot\phi^*(x)$. $\int \phi_k(x)\phi_l^*(x)dx = \delta_{kl}$ because the basis is othonormal. Finally, we obtain a simple formula:

$$Q(f,g) = Re\left(\sum_k f_k \cdot g_k \cdot \rho_k\right) \tag{2}$$

Value $Q(f,g)$ is a scalar product of two functions $Q = \langle f,g\rangle$. We can normalise our measure and define the correlation coefficient:

$$r = \frac{Re\left(\left\langle \widetilde{f}, \widetilde{g}\right\rangle\right)}{\sqrt{\left(\left\langle \widetilde{f}, \widetilde{f}\right\rangle \cdot \langle \widetilde{g}, \widetilde{g}\rangle\right)}} \tag{3}$$

where $\widetilde{f_k} = f_k - Mean(f)$; $\widetilde{g_k} = g_k - Mean(g)$. The correlation coefficient $r \in [-1,1]$.

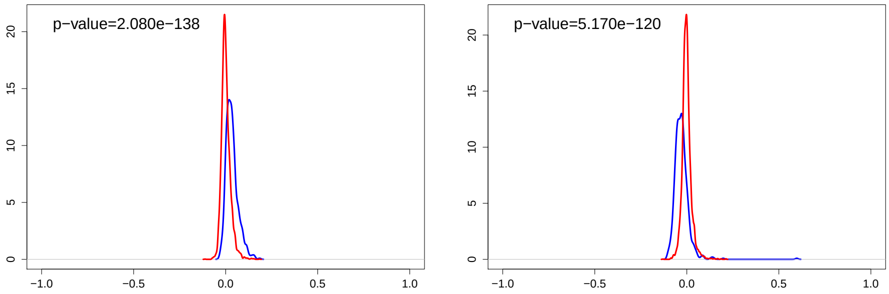
Permutation test

As FFT allows very quick integration of kerneled convolutions, we assess the statistical significance by permutations. As a null model representing the behaviour in the absence of a dependence, different models can be used, depending on our expectations about the spatial properties of features. The most common null distribution is obtained by setting the corresponding between the first feature in a window and the second feature in a randomly chosen window. As a result of the test, we obtain two distributions of correlations of the two analyzed features. The first is for all the (non-overlapping) windows in the investigated region (e.g. in a chromosome or in the whole genome), and is referred to as the real data. The second distribution is obtained by permutations; e.g., we relate random pairs of windows and correlate the first feature form the first window and the second feature from the second window. The main test we apply on these two distributions is the Wilcoxon that tests whether the medians of the distributions are the same. If the median of the real data distribution is significantly higher than that for permutations, the two sets of features are positively correlated. If it is significantly lower, negative correlation is detected. Additionally, the distributions contain a lot information about the relations of the features: we can compare the shapes of the distributions, analyse the number of modes, the weight of tails, etc, thus detecting the presence of genome regions that provide unusual values of correlation.

Results

To test the suggested method, we analysed the interrelation between the histone modification marks that are known to be the marks of eu- and hetero-chromatin, and the relation between two of them (one for each type) and the mRNA-seq signal. All the data were taken from Human Epigenome Atlas (`http://www.genboree.org/epigenomeatlas`) for Fetal Brain tissue.

Fig. 1 shows the result of our method when applied to three features: H3K4me1 (euchromatine histone modification), H3K27me3 (heterochromatine histone modification), and mRNA-Seq. As expected (see, e.g., [24]), H3K4me1 is positively correlated with mRNA, and negatively correlated with H3K27me3. The results are in concordance with a similar test for correlation between highly transcribed gene promoters and these marks [1].



(a)

(b)

Fig. 1: Density for correlation function (a) for histone modification *H3K4me1* and *mRNA-Seq* (b) for histone modification *H3K27me3* and *mRNA-Seq*. P-value is calculated using Wilcoxon test. Red — background distribution; Blue — observed distribution.

Fig. 2 shows the results of a consistency test on the histone marks. We tested the correlation of an accetylation euchromatine ('active') mark with two methylation 'active' marks, and the correlation of the accetylation 'active' mark with a methylation heterochromatine ('repressive') mark. As it could be expected, the first two tests revealed a positive correlation, while the last test revealed a negative correlation.

Conclusion

We have developed a novel method for fast assessment of the correlation between pairs of genomic features and of its statistical significance. The method is based

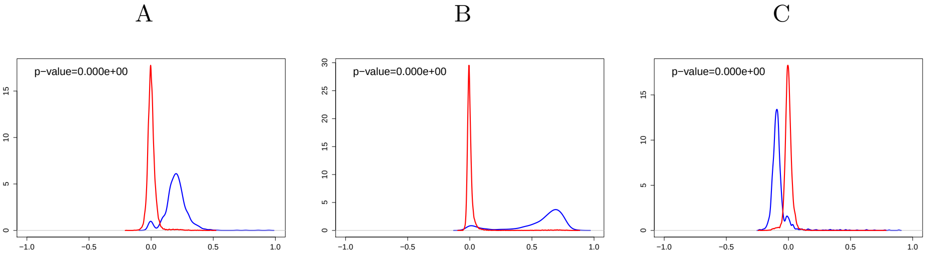


Fig. 2: Distributions of correlation coefficients that are obtained in histone mark consistency test. Red — background distribution; Blue — observed distribution. A: distribution for the 'active' marks (H3K9ac,H3K4me1); B: Extreme case correlation distribution for 'active' marks (H3K9ac,H3K4me3). C: Distributions for correlation of 'active' and 'repressive' marks (H3K9ac,H3K9me3).

on FFT, and thus is extremely fast. Therefore, it can be run on a set of windows, and the resulting distribution can be compared with the similar one that was obtained for permuted windows. The relation between the two distributions can be assessed by any statistical routine, thus providing information about the correlation of the two features along the genome. We checked the test on a real dataset, and found that the results that are provided by the test are in correspondence with our knowledge about the data.

References

1. Favorov, A., Mularoni, L., Cope, L.M., Medvedeva, Y., Mironov, A.A., Makeev, V.J., Wheelan, S.J.: Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Comput Biol **8**(5) (May 2012) e1002529
2. Ramsey, S.A., Knijnenburg, T.A., Kennedy, K.A., Zak, D.E., Gilchrist, M., Gold, E.S., Johnson, C.D., Lampano, A.E., Litvak, V., Navarro, G., Stolyar, T., Aderem, A., Shmulevich, I.: Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. Bioinformatics (Oxford, England) **26**(17) (September 2010) 2071–2075 PMID: 20663846.
3. Bickel, P.J., Boley, N., Brown, J.B., Huang, H., Zhang, N.R.: Subsampling methods for genomic inference. The Annals of Applied Statistics **4**(4) (December 2010) 1660–1697 Zentralblatt MATH identifier: 05910045; Mathematical Reviews number (MathSciNet): MR2829932.
4. Bickel, P.J., Brown, J.B., Huang, H., Li, Q.: An overview of recent developments in genomics and associated statistical methods. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences **367**(1906) (November 2009) 4313–4337 PMID: 19805447.
5. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer,

T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D., Kent, W.J.: The UCSC genome browser database: extensions and updates 2013. Nucleic acids research **41**(D1) (January 2013) D64–69 PMID: 23155063.

6. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., Thomson, J.A.: The NIH roadmap epigenomics mapping consortium. Nature Biotechnology **28**(10) (2010) 1045–1048
7. ENCODE Project Consortium: The ENCODE (ENCyclopedia of DNA elements) project. Science (New York, N.Y.) **306**(5696) (October 2004) 636–640 PMID: 15499007.
8. Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., Wang, T., Zhong, S.: Comparative epigenomic annotation of regulatory DNA. Cell **149**(6) (June 2012) 1381–1392 PMID: 22682255.
9. Arvey, A., Agius, P., Noble, W.S., Leslie, C.: Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome research **22**(9) (September 2012) 1723–1734 PMID: 22955984.
10. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., Sim, H.S., Peh, S.Q., Mulawadi, F.H., Ong, C.T., Orlov, Y.L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K.I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M.J., Cheung, E., Liu, E., Sung, W.K., Snyder, M., Ruan, Y.: Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell **148**(1-2) (January 2012) 84–98 PMID: 22265404.
11. Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E.H., Farnham, P.J., Jin, V.X.: Integration of hi-c and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic acids research **40**(16) (September 2012) 7690–7704 PMID: 22675074.
12. Toedling, J., Ciaudo, C., Voinnet, O., Heard, E., Barillot, E.: Girafe–an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. Bioinformatics (Oxford, England) **26**(22) (November 2010) 2902–2903 PMID: 20861030.
13. Philipp, E.E.R., Kraemer, L., Mountfort, D., Schilhabel, M., Schreiber, S., Rosenstiel, P.: The transcriptome analysis and comparison explorer–t-ACE: a platform-independent, graphical tool to process large RNAseq datasets of non-model organisms. Bioinformatics (Oxford, England) **28**(6) (March 2012) 777–783 PMID: 22285826.
14. Halachev, K., Bast, H., Albrecht, F., Lengauer, T., Bock, C.: EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome biology **13**(10) (October 2012) R96 PMID: 23034089.
15. Zackay, A., Steinhoff, C.: MethVisual - visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing. BMC research notes **3** (2010) 337 PMID: 21159174.
16. Giles, K.E., Gowher, H., Ghirlando, R., Jin, C., Felsenfeld, G.: Chromatin boundaries, insulators, and long-range interactions in the nucleus. Cold Spring Harbor symposia on quantitative biology **75** (2010) 79–85 PMID: 21047907.
17. Chikina, M.D., Troyanskaya, O.G.: An effective statistical evaluation of ChIPseq dataset similarity. Bioinformatics (Oxford, England) **28**(5) (March 2012) 607–613 PMID: 22262674.

18. Zhang, Z.D., Paccanaro, A., Fu, Y., Weissman, S., Weng, Z., Chang, J., Snyder, M., Gerstein, M.B.: Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome research **17**(6) (June 2007) 787–797 PMID: 17567997.
19. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., Loh, Y.H., Yeo, H.C., Yeo, Z.X., Narang, V., Govindarajan, K.R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.K., Clarke, N.D., Wei, C.L., Ng, H.H.: Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell **133**(6) (June 2008) 1106–1117 PMID: 18555785.
20. Carstensen, L., Sandelin, A., Winther, O., Hansen, N.R.: Multivariate hawkes process models of the occurrence of regulatory elements. BMC bioinformatics **11** (2010) 456 PMID: 20828413.
21. Huen, D.S., Russell, S.: On the use of resampling tests for evaluating statistical significance of binding-site co-occurrence. BMC bioinformatics **11** (2010) 359 PMID: 20591178.
22. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., Zhao, K.: Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics **40**(7) (July 2008) 897–903 PMID: 18552846.
23. Fu, A.Q., Adryan, B.: Scoring overlapping and adjacent signals from genome-wide ChIP and DamID assays. Molecular bioSystems **5**(12) (December 2009) 1429–1438 PMID: 19763325.
24. Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guig, R., Birney, E., Weng, Z.: Modeling gene expression using chromatin features in various cellular contexts. Genome biology **13**(9) (2012) R53 PMID: 22950368.