Towards better prioritization of epigenetically modified DNA regions

Ernesto Iacucci[1], Dusan Popovic[1], Georgios A. Pavlopoulos[1], Léon-Charles Tranchevent[1], Marijke Bauters[2], Bart De Moor[1], Yves Moreau[1]

[1]ESAT-SCD / IBBT-K.U.Leuven Future Health Department, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium
[2]Department of Human Genetics / IBBT-K.U.Leuven Future Health Department, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium

{Ernesto.Iacucci, Dusan.Popovic, Georgios.Pavlopoulos, Leon-Charles.Tranchevent, Bart.DeMoor, Yves.Moreau}@esat.kuleuven.be

{marijke.bauters}@cme.vib-kuleuven.be.be

Abstract: Epigenetic modifications of the genome can cause profound changes in phenotype of an organism. Experimental methods allow us to detect regions of the DNA that have been epigenetically modified; these regions are said to be *enriched* in a queried state versus a control. Detecting the enriched regions is not a simple matter as making sense of the data involves multiple analytical steps and often results in false calls. In this study, we analyze the utility of using additional features of the data (such as the transcription start site (TSS) and the histone coverage) to detect enrichment. We train a decision tree ensemble using these three features and review how well they identify regions that are truly enriched (as validated by q-PCR). We find that the enrichment score derived directly from ChIP-chip experiment data is less informative than the histone coverage.

Keywords: ChIP-chip, data integration, protein-DNA, machine learning, decision trees

1 Introduction

The detection of protein-DNA interactions is an important area of research. Protein-DNA interactions account for various cellular events such as DNA repair and transcription factor binding [1-3]. Transcription factors regulate the expression level of gene products that carry out the majority of processes in the cell. Histone-DNA interactions are a specific type of protein-DNA interactions that also influence the expression of genes. Indeed, DNA that is wound around histone-bodies (the complex form of histones) is less accessible to the cellular transcriptional machinery and thus genes

located in these regions are less likely to be expressed [4,5]. These modifications are considered epigenetic as they alter the expression of genes while not changing their sequences.

A widely used technique to measure protein-DNA interaction is chromatin immunoprecipitation followed by DNA microarray hybridization (ChIP-chip). Using ChIP-chip, one is able to identify areas of the genome that are enriched between two conditions of interest (e.g., disease vs. control) [1,6]. Detecting the enriched regions is not a simple matter as making sense of the data involves multiple analytical steps and often results in false calls [7,8]. In this study, we assess whether using additional features enhance the detection of enriched regions [9,10]. In addition to the enrichment scores, extracted from ChIP-chip data, the transcription start site (TSS) and histone coverage scores are defined and used to train a decision tree based algorithms.

While the primary feature resulting from a ChIP-chip experiment is the enrichment score for a region, the other two features are easily derived. The TSS score is the distance of the region to the nearest predicted TSS. The histone coverage is a unit value which is calculated from a regions size (in base-pairs) in relation to the size of a full turn of the DNA around a histone body (147 base-pairs).

We then review how well these three features perform in predicting the regions that are truly enriched (as validated by q-PCR).

2 Methods

Our dataset consists of 25 DNA regions for which we have ChIP-chip enrichment scores, region sizes and distances to the nearest transcription start site, and validated q-PCR values. Our dataset is derived from ChIP-chip experiments essaying fragile-X patient samples (data unpublished). The q-PCR values define the positive and negative examples and will be considered binary for the purposes of this work.

- The ChIP-chip *enrichment score* is derived from a data analysis procedure described in [11,12]. Briefly, the data is processed as follows:
- The outliers in the data are removed (probes in a 5 probe window are averaged and probes which are over 2 standard deviation from the mean are removed).
- The data is normalized, by adjusting the mean of entire distribution to zero.
- The differences between the two samples are calculated (one sample is the condition/disease sample and the other would be the control).
- The data is smoothed (a 3-point moving average is calculated for each peak).
- The probes, which show significant differences, are identified (those over 2 standard deviations from the mean).
- The regions of consistent difference defined by multiple probes (4 probes of a 5 probe window) are called (flagged as significant).

The *transcription start site* (*TSS*) *feature* is calculated as the distance from the nearest TSS. These distances (measured in base-pairs), are then mapped to an integer score which varied from 0 to 5. The *histone coverage* is a feature which is computed from the size (measured in base-pairs) of the enriched region. The size of the region is transformed into a unit value by applying the equation displayed in Figure 1.

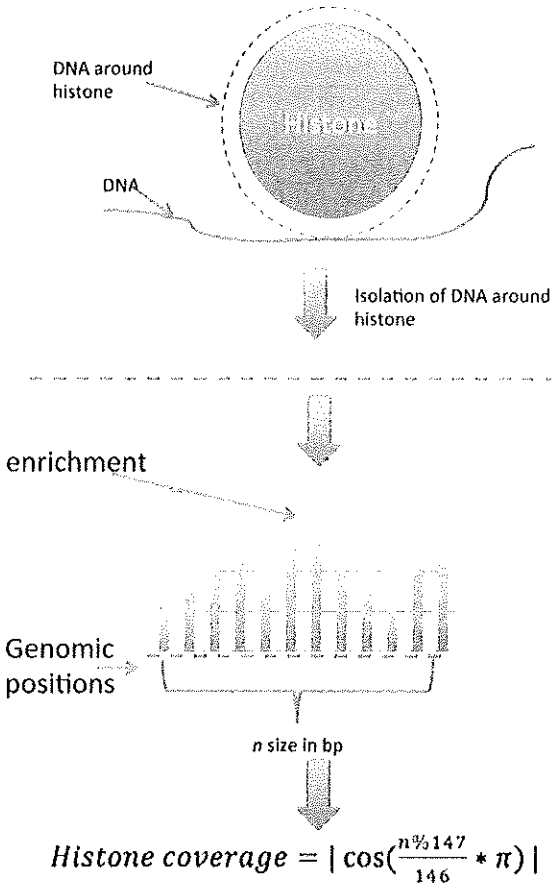

Fig. 1. Histone coverage score calculation

The dataset, consisting of the three features and the validated q-PCR outcomes, is then feed to a decision tree-learning algorithm (classregtree, Matlab v7.10.0). In addition, a bagged decision tree ensemble classifier on the whole dataset is also trained.

This algorithm builds individual trees on the bootstrap replicates of the original dataset and then uses out-of-bag observations to compute unbiased estimates of the classification error. This is often exploited to measure feature importance. For each fea-

ture, its values across all the observations are permuted, after which the difference in mean squared error is examined. Eventually, a higher positive difference implies greater importance for that feature. Furthermore, we validated our results using leave-one-out cross validation.

3 Results and Discussion

We run a decision tree learning algorithm on the whole dataset to examine which features are selected as the most informative and in which order. We construct a ROC curve for each feature and examine the AUC as a heuristic to determine which features are the most important. Out of the three features considered, we find that enrichment performs the poorly (AUC TSS: 0.62, AUC Enrichment Score 0.60, AUC Histone score: 0.73). This result suggests that the use of enrichment scores alone is not an optimal strategy to predict truly enriched regions. We observe that the TSS score and histone coverage, are necessary to improve performance in the prediction task. This observation is consistent with the results from the out-of-the-bag feature importance analysis (see Figure 2).

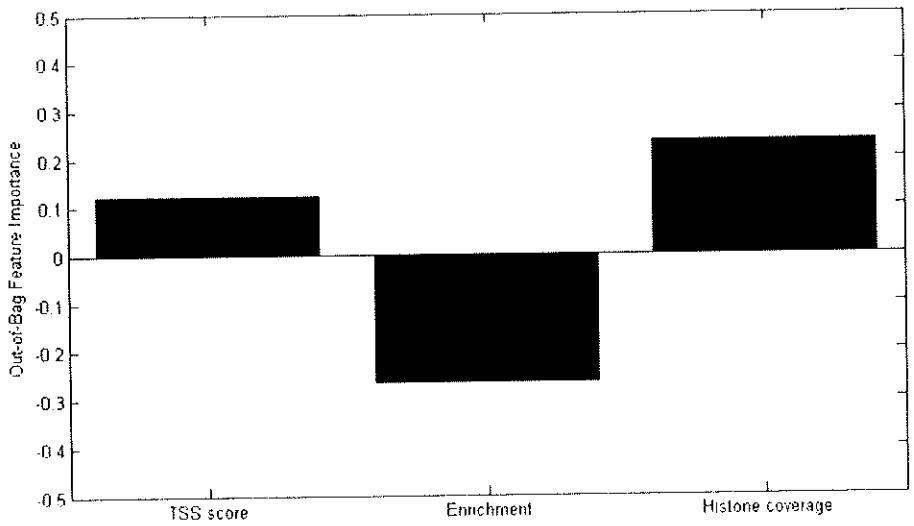

Fig. 2. Out of bag importance of features

Figure 2 demonstrates that the highest positive difference occurs with the histone coverage, which implies greater importance of this feature. Surprisingly, the enrichment feature is associated to a negative difference, indicating that it is the least important of the features. In order to illustrate this finding with the original data, we

create a scatter plot that compares the histone coverage with the enrichment value while at the same time indicating the positive and negative regions (see Figure 3).

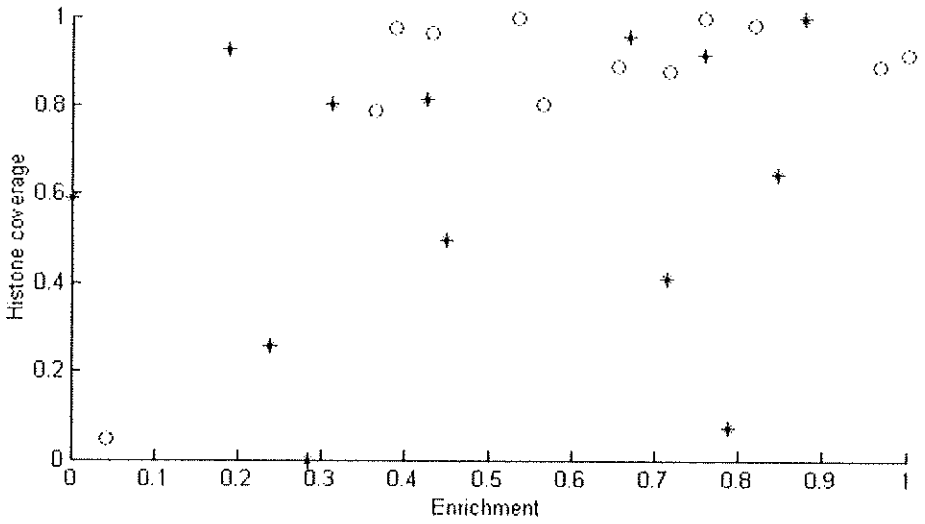

Fig. 3. Scatter plot of histone coverage vs enrichment score. Circles indicate positive examples and crosses indicate negative examples

Figure 3 demonstrates that positive examples (circles) are concentrated at higher histone coverage values while they are spread across high and low enrichment values. Negative examples (crosses) are also spread across high and low enrichment values but are mostly found at lower histone coverage values.

The utility of the TSS score and the histone coverage is more apparent when one considers that the decision tree constructed using the whole dataset has a topology which determines the first split on the histone coverage and the second split on the TSS score and determines no splits on the enrichment score (see Figure 4).

In order to assess the reliability of this approach, we ran 100 iterations of a leave-one-out cross validation analysis. The results were as follows: using the enrichment feature alone, the random forest algorithm has a mean performance (accuracy) of 0.44 (st. dev. 0.022), when we use all three features the value rises to 0.64 (st. dev. 0.045), when we use the TSS score and the Histone coverage (and no enrichment value), the best value, of 0.76 (st. dev. 0.022), is achieved.

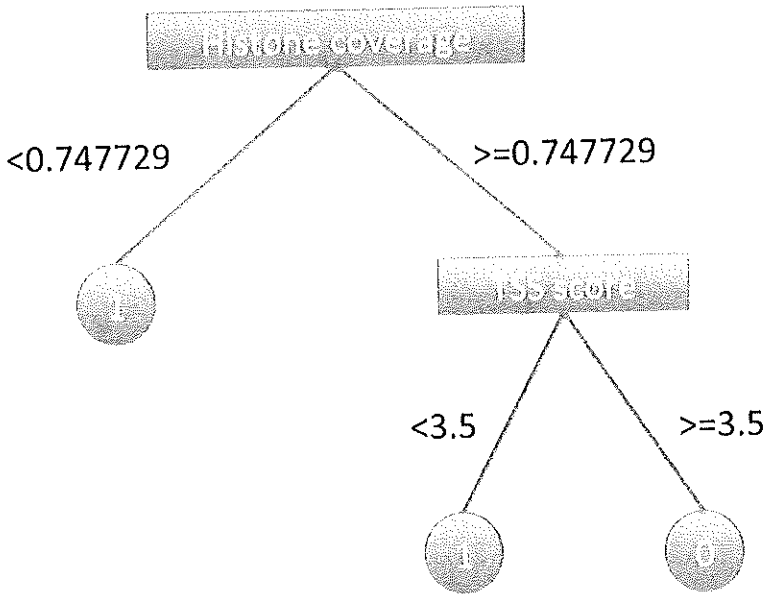

Fig. 4. Decision Tree

As presented in Table 1, the results are consistent with the ones obtained single decision trees.

Table 1. Metric comparison across 100 experiments

Features	Decision trees		Random forest	
	mean	*std.*	*mean*	*std.*
TSS score, Histone coverage	0.6380	0.0237	0.7600	0.0223
TSS score, Histone coverage, Enrichment	0.5948	0.0315	0.6400	0.0446
Enrichment	0.4508	0.0196	0.4400	0.0223

4 Conclusions

This work present novel insight into the task of predicting true enrichment in regions detected by ChIP-chip experimentation. Our main technical contribution is two-fold. First, we demonstrate that the use of enrichment scores alone is not an optimal strategy. Second, we show that the use of two additional features, namely TSS and histone coverage, provide unique information, and are necessary to improve the prediction results. Looking forward, we plan to examine the integration of other features and the development of other strategies, which might increase predictive power.

5 Acknowledgements

Funding: The authors would like to acknowledge support from: Research Council KUL:ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys , START 1, several PhD/postdoc & fellow grants. Flemish Government: FWO: PhD/postdoc grants, projects G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR) IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3 FOD:Cancer plans, IBBT. Bel- gian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007- 2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH; CHeartED

6 Reference List

1. Bieda, M., Xu, X., Singer, M.A., Green, R., Farnham, P.J.: Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res **16**(5), 595-605 (2006)
2. Kreuz, M., Rosolowski, M., Berger, H., Schwaenen, C., Wessendorf, S., Loeffler, M., Hasenclever, D.: Development and implementation of an analysis tool for array-based comparative genomic hybridization. Methods Inf Med **46**(5), 608-613 (2007)
3. Pelizzola, M., Koga, Y., Urban, A.E., Krauthammer, M., Weissman, S., Halaban, R., Molinaro, A.M.: MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome Res **18**(10), 1652-1659 (2008)
4. Dowell, R.D.: Transcription factor binding variation in the evolution of gene regulation. Trends Genet **26**(11), 468-475 (2010)
5. Gilchrist, D.A., Fargo, D.C., Adelman, K.: Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal

widespread regulation of transcription elongation. Methods **48**(4), 398-408 (2009)
6. MacQuarrie, K.L., Fong, A.P., Morse, R.H., Tapscott, S.J.: Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet **27**(4), 141-148 (2011)
7. Toedling, J., Huber, W.: Analyzing ChIP-chip data using bioconductor. PLoS Comput Biol **4**(11), e1000227 (2008)
8. Toedling, J., Skylar, O., Krueger, T., Fischer, J.J., Sperling, S., Huber, W.: Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics **8**, 221 (2007)
9. Chen, K.B., Zhang, Y.: A varying threshold method for ChIP peak-calling using multiple sources of information. Bioinformatics **26**(18), i504-510 (2010)
10. Johnson, D.S., Li, W., Gordon, D.B., Bhattacharjee, A., Curry, B., Ghosh, J., Brizuela, L., Carroll, J.S., Brown, M., Flicek, P., Koch, C.M., Dunham, I., Bieda, M., Xu, X., Farnham, P.J., Kapranov, P., Nix, D.A., Gingeras, T.R., Zhang, X., Holster, H., Jiang, N., Green, R.D., Song, J.S., McCuine, S.A., Anton, E., Nguyen, L., Trinklein, N.D., Ye, Z., Ching, K., Hawkins, D., Ren, B., Scacheri, P.C., Rozowsky, J., Karpikov, A., Euskirchen, G., Weissman, S., Gerstein, M., Snyder, M., Yang, A., Moqtaderi, Z., Hirsch, H., Shulha, H.P., Fu, Y., Weng, Z., Struhl, K., Myers, R.M., Lieb, J.D., Liu, X.S.: Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res **18**(3), 393-403 (2008)
11. Sharp, A.J., Migliavacca, E., Dupre, Y., Stathaki, E., Sailani, M.R., Baumer, A., Schinzel, A., Mackay, D.J., Robinson, D.O., Cobellis, G., Cobellis, L., Brunner, H.G., Steiner, B., Antonarakis, S.E.: Methylation profiling in individuals with uniparental disomy identifies novel differentially methylated regions on chromosome 15. Genome Res **20**(9), 1271-1278 (2010)
12. Sharp, A.J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S.B., Dupre, Y., Antonarakis, S.E.: DNA methylation profiles of human active and inactive X chromosomes. Genome Res **21**(10), 1592-1600 (2011)