Beyond Enrichment Scores

Novel Strategies for Predicting Epigenetically Modified Regions

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

[1] ESAT-SCD / IBBT-KU Leuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium

[2]Department of Human Genetics / IBBT-KU Leuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium

[*]Ernesto.Iacucci@esat.kuleuven.be

Abstract

Epigenetic modifications of the genome are the key regulatory agents in the grand scheme of genomics. These modifications can cause profound changes in the phenotype of an organism. Detecting these modified regions is not a simple matter as making sense of the data involves several analytical steps. Another complication in the workflow is the occurrence of false calls made in the data analysis step. It is therefore an imperative to accurately prioritize the potentially modified regions for the final stage of time consuming and costly wet-lab validation.

In this study, we analyze the utility of data fusion as it applies to existing and novel features (such as sequence conservation and histone coverage) to improve the prioritization of candidate modified regions. We train several machine learning techniques to determine which is the most appropriate for the task. We find that the Random Forest method performs best with a precision of 0.70.

Surprisingly, we find that the enrichment score derived directly from ChIP-chip experiment data is less informative other features such as the histone coverage. In addition, we test the ranking of new qPCR validated examples and find that we achieve an AUC of 0.62.

Keywords

Epigenetics; ChIP-chip; Machine Learning; Prioritization

Introduction

Epigenetic modifications of the genome are the key regulatory agents in the grand scheme of genomics. Changes in the epigenetic state influences various cellular events such as DNA repair and transcription factor binding [Bieda et al., Kreuz et al., Pelizzola et al.]. The epigenetic state of an organism can influence the accessibility of the DNA and thus the transcription of regions in the genome. The accessibility of the DNA is mediated by a protein-DNA interaction called histone wrapping. Briefly described, this phenomenone holds that DNA which 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 [Dowell, Gilchrist et al.]. Thus, the detection of protein-DNA interactions is a key area of study in epigenetics. One experimental technique used to measure protein-DNA interaction is chromatin immuno-precipitation followed by DNA microarray hybridization (ChIP-chip). Using ChIP-chip, one is able to identify areas of the genome that are enriched (modified) between two conditions of interest (e.g., disease vs. control) [Bieda et al., MacQuarrie et al].

The modification of DNA through methylation is a well studied phenomena which can cause profound changes in the phenotype of an organism. In order to better characterize the pattern and effect of methylation, a number of methods have been put forth. Two such packages which employ popular strategies for the analysis of this data are Ringo [Toedling et al.] and MEDME [Pelizzola et al.]. The simplest approach to this analysis is taken in the Ringo package as it determines an enrichment score through direct comparison of two-color microarray intensities. More complex is the concept behind the modeling experimental data with MeDIP enrichment (MEDME) approach which is to estimate DNA histone methylation through a proposed combination of experimental and analytical methodology.

While these methods make straight-forward use of enrichment scores, they do not attempt to incorporate a range of features which may be informative in this prediction task. Here we seek to examine these features and the various machine learning methods available to researchers. Through benchmarking we find the optimal combination of machine learning methods and features to apply to this learning task.

While many features exist which describe the candidate DNA regions, we turn our attention to the most ubiquitous and well currated features available for this study. In addition to the enrichment score (described above), we consider the distance to the nearest transcription start site (TSS), the sequence conservation, the associated expression level, and the histone score. The distance to transcription start site represents information related to the spatial arrangement of the candidate regions in relation to gene transcription initiation events. The expression level is a measure which also represents information related to expression level of the nearest gene in the genome. The sequence conservation feature represents the importance of the region as it may be conserved through evolution in order to maintain function. The histone score (described below) is a novel score which we use in order to relate the size of the region to it's wrapping around the histone bodies. Clearly these features represent disparate information which we

consider can be used, through data fusion, in order to better classify regions as modified (enriched) or not-modified (not-enriched).

Computational prediction allows for facilitation of high-dimensional problems which may contain complex and enigmatically related data. The high through-put era of computational biology has added further difficulty as the problems inherent to this field often involve thousands of data points which cannot be processed without computational resources. The prediction of epigenetically modified regions is one such problem for which we apply methods for the prediction of candidate regions.

The study of using features to predict classes (such as "modified" or "not-modified") is an active area of research in the domain of machine learning. Generally, the process of learning from the features to predict classes, in a "supervised" context, is a process which starts with partitioning the data. Data is partitioned into training and test data. Training data is used by the classifiers to construct rules based on known feature measurements of members of the classes. Once a classifier is built, its performance can be assessed using the test data.

Below, we analyze the utility of data fusion as it applies to existing and novel features (such as sequence conservation and histone coverage) to improve the prediction of candidate modified regions. We assay several machine learning techniques to determine which one gives the best performance. In addition, we assess the level of information present in the different features and determine which are most appropriate for use in this task. Finally, we present our findings and draw conclusion based on them in order to provide guidance for future work in this area.

Methods

Features and Training Examples

The sequence conservation scores were downloaded from the UCSC genome browser on April 2012 using Galaxy. More precisely, the phastConsElements46wayPrimates table from the conservation track was used to retrieve the conservation scores between the human genome and various primate genomes.

The transcription start sites have been downloaded on March 2012 from EnsEMBL using BioMart. The complete file contains 190 243 transcripts and their TSS, of which 7558 are located on chromosome X. The transcription start site (TSS) feature, as we use it in analysis, is calculated as the distance between start of the enriched region and the nearest TSS expressed in base pairs. The missing values are estimated as the mean value of TSS distance of regions where the later is available.

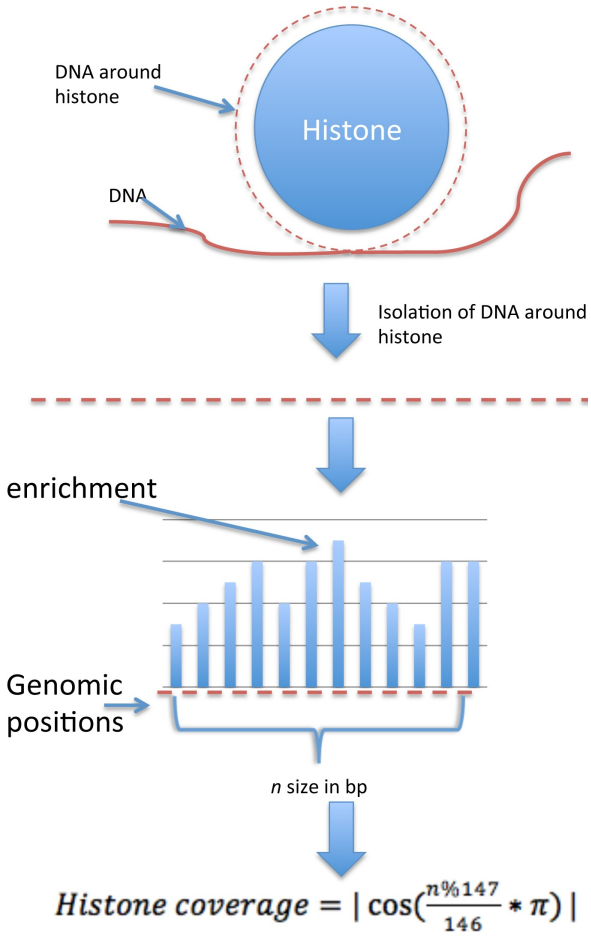


FIG. 1 CREATION OF THE HISTONE COVERAGE SCORE

The expression profile for the candidates was taken from the GNF human expression atlas. The data was normalized (values were mean-zeroed and the standard deviation was set to one) [Su et al.] and the missing values are replaced by mean. The enrichment score (log2scores) was derived as described in [Sharp et al., Sharp et al.] using the ChIP-chip data described below.

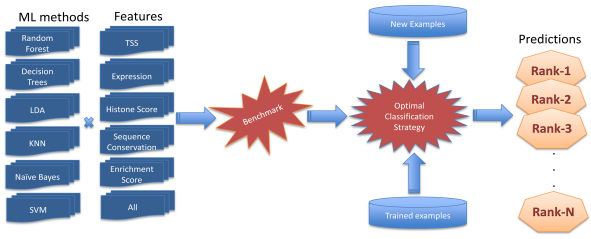


FIG. 2 WORKFLOW OF BENCHMARK

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. All of the features are scaled to (0,1) interval.

Outcomes for the training examples were derived from the UCSC Genome Browser (NCBI36/hg18 Assembly), using the ENCODE Histone Modifications from the Broad Institute, generated via ChIP-seq, and more specifically those displayed for the

lymphoblastoid cell line GM12878 (ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me3, GM12878).

ChIP-chip and qPCR Validation

Epstein-Barr virus-transformed peripheral blood lymphocytes (EBV-PBL) of four healthy individuals were cultured according to standard tissue culture procedures. Ten million cells were used for each chromatin immunoprecipitation (ChIP) experiment. ChIPs with anti-H3K4me3 (pAb-003-050, Diagenode) were performed as described in Frank et al. [Frank et al.] with some minor modifications. ChIP and input DNA were amplified using the GenomePlex® Complete Whole Genome Amplification kit (Sigma) according to manufacturer's protocol. Amplified DNA was sent to the NimbleGen Service (Iceland) and hybridized to a custom X-chromosome specific array. Confirmation experiments were done via realtime qPCR on the LightCycler® 480 (Roche) using 10 nanograms of both immunoprecipitated and input DNA.

Design of the Classification Benchmark

In the context of machine learning and statistics, the classification is a problem of assigning previously unseen observation to one of the fixed number of predefined classes. For doing this, one typically constructs discriminatory function, called classifier, using training data and some suitable algorithm. The choice of later determines what kind of function can be inferred, how well certain properties of the data will be handled (size, dimensionality, distributional assumptions) and how will classifier generalize on new examples. As all of encountered here is problem-dependent, and as there are no algorithm that outperform all other in all situations [Wolpert et al.], one typically tries several approaches for a given problem.

In order to assess performance of classifiers build using different combinations of features and algorithms, we conduct extensive benchmarking of our dataset. During each and every iteration of the benchmark a bootstrap sample [Efron] is taken out for the training with six different methods – Decision trees, Random forests, K-nearest neighbours, Linear discriminant analysis, Naïve Bayes and Support vector machines. In addition, these have been trained using either all features or on subsets created by removing one feature at the time; resulting in 36 classifiers in total. Subsequently, left-out examples from the original data set (typically 36.8%) have been used for testing. This procedure has been repeated in randomised manner for hundred times to estimate stability of the result (see figure #2).

Finally, given the best methods, as determined during benchmarking phase, we conduct a qPCR validation of 25 of our predictions in order to determine the utility of our approach.

Classification Algorithms

Our choice of classification algorithms to be included in the benchmark has been motivated by the two main rationales - we tried to cover variety of fundamentally different methods that also have been already proven their merit in the field of bioinformatics [Larranaga et al.].

Decision trees [Breiman] is a class of simple algorithms for regression and classification which are based on greedy, top-down, recursive partitioning of the feature space. Trained classification tree predicts membership of the new instance given values of one or several of its attributes sequentially, by testing them against rules embedded in nodes of previously learned hierarchical structure.

Random forest [Breiman] belongs to group of ensemble methods. Essentially, it is collection of decision trees built on different bootstrap samples extracted from training data. An additional randomness is injected to algorithm by limiting the choice of candidate variables for each split in each decision tree to randomly chosen subset, typically of size proportional to logarithm or square root of total number of variables. The final classification is reached by majority vote among trees in ensemble.

K-Nearest Neighbours [Bruckstein et al.] is a simple instance-based learning method that classifies previously unseen examples using labels of k closest instances from the training set. Here k stands for number of neighbours that are considered, and it is user-defined parameter usually optimized by cross-validation.

Linear Discriminant Analysis [Bruckstein et al., Hand] is the method that seeks for reduction of dimensionality of the original classification problem by projecting observations to the single line such that most of class discriminatory information is preserved. This line is chosen to maximize the ratio of between-class to within-class variance. The method requires Gaussian distribution of the classes and it is very sensitive on deviations from this assumption.

Naïve Bayes [Bayes] is simple probabilistic method that strongly assumes independence of the features given the class variable, which renders it less sensitive to the curse of dimensionality. In addition, it has been shown both in practice and theoretically [Hand, Zhang et al.] that Naïve Bayes can produce reliable classification even in some cases where independence assumption is considerably violated.

Support Vector Machines [Vapnik, Vapnik] are class of methods for nonlinear classification and regression based on the statistical learning theory and structural risk minimization. Using a kernel function they map observations to new, often infinite dimensional, feature space where optimal (margin-maximising) separating hiperplane is then constructed. Due to this, the support vector machines are able of inferring arbitrary

shaped decision boundaries and of handling highly dimensional data.

For all classification algorithms we use their respective Matlab 7.10 implementations – classregtree class for the decision trees, TreeBagger class for the random forests, function knnclassify for the KNN, function classify (with argument "linear") for the LDA, class NaiveBayes for Naïve Bayes classification and functions svmtrain and svmclassify for the support vector machines. The most of the used functions and classes are part of the Statistical Toolbox, except that for KNN and SVMs (Bioinformatics Toolbox).

Results and Discussion

Our benchmark revealed several key findings. Significant among the findings were the answers to our initial questions concerning which were the best choice of classifier and features. Below, we detail our findings and discuss their significance.

Features

As described above, this work represents a comprehensive view of data fusion as it applied for the prediction of candidate modified regions. As such, we turn our attention to the various drops in performance in figure #3. Particularly, we see the drop in performance across all classifiers seems to consistently occur when the TSS feature is removed from the various classifiers. This can be explained by the fact that the spatial component of the epigenetic modifications plays an important role in regulation. In addition we find that there is generally an increase in performance when the enrichment score is removed from the analysis.

Classifiers

The classifier which provides the best accuracy is the random forest algorithm with a score of 0.70 (see figure #3). This algorithm builds individual trees on the bootstrap replicates of the original dataset. In our setting, the classifiers were trained with 120 examples and thus the performance of this classifier may increase if the number of training examples were to increase. Other classifiers also presented notable results, such as the Naïve Bayes classifer (with the TSS removed) which has a sensitivity of 0.75, this result is notable but should be considered in combination with the low score of 0.32 for specifically, thus suggesting that a low threshold is responsible for these result. Indeed, the precision of this classifier reflects this at 0.55.

Prioritization

In order to determine the utility of random forest classifier for the purpose of prioritization, we assess the AUC of the 25 predictions which were validated using qPCR. While the resulting value of 0.62 (see figure #4) was not extremely high, this may be attributed to the fact that the classifier would improve in performance with more training examples. We consider that this is an area for further research in this field and are encouraged by the competitive accuracy value obtaining in the training set.

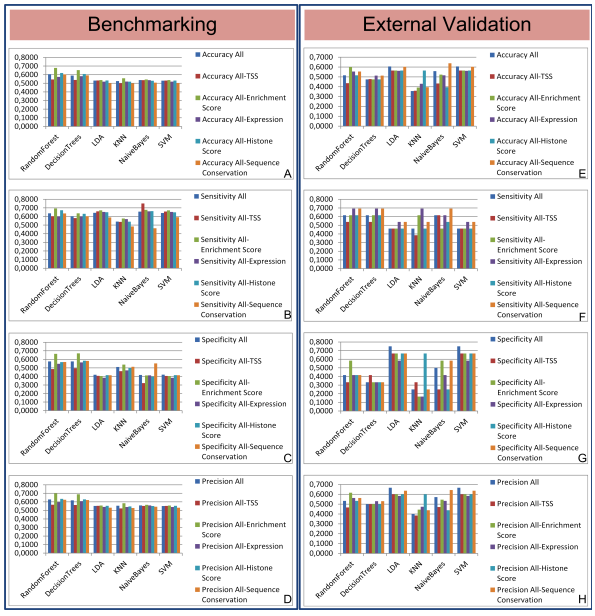


FIG. 3 BENCHMARK AND EXTERNAL VALIDATION SET

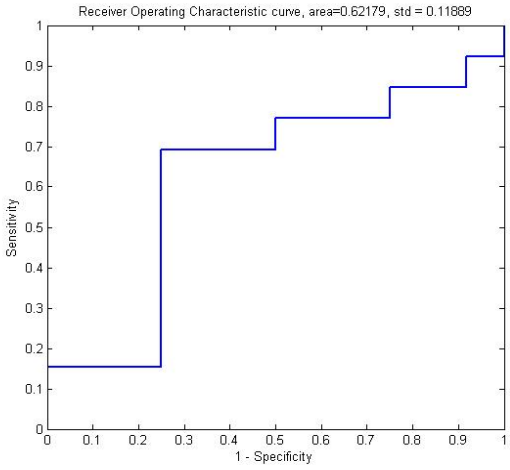


FIG. 4 ROC OF NEW EXAMPLES

Conclusions

The results from this work suggest several key findings. First, there exists a major advantage to using a wide range of features versus the tradition enrichment score. Second, we introduce the use of a novel feature, histone coverage, which appears to contribute the highest amount information to the prediction task while the enrichment score is perhaps the worst feature. Finally, as the best performing machine learning method was the random forest algorithm which provides a high quality prioritized list of candidates which can easily be used for wet-lab validation.

ACKNOWLEDGMENT

Funding: the authors would like to acknowledge support from the Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys , START 1), FWO (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 (Silicos, SBO-BioFrame, SBO-MoKa, TBM-IOTA3), FOD (Cancer plans), IBBT, Belgian 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).

REFERENCES

Bayes, T. "An essay towards solving a problem in the doctrine of chances." 1763. MD Comput 1991, 8:157-171.

Bieda, M, Xu, X, Singer, MA, Green, R, and Farnham, PJ. "Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome." Genome Res 2006, 16:595-605.

Breiman, L. "Prediction games and arcing algorithms." Neural Comput 1999, 11:1493-1517.

Bruckstein, AM, Cover, TM. "Monotonicity of linear separability under translation." IEEE Trans Pattern Anal Mach Intell 1985, 7:355-358.

Dowell, RD. "Transcription factor binding variation in the evolution of gene regulation." Trends Genet 2010, 26:468-475.

Efron, B. "The bootstrap and Markov-Chain Monte Carlo." J Biopharm Stat 2011, 21:1052-1062.

Frank, SR, Schroeder, M, Fernandez, P, Taubert, S, and Amati, B. "Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation." Genes Dev 2001, 15:2069-2082.

Gilchrist, DA, Fargo, DC, and 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 2009, 48:398-408.

Hand, DJ. "Statistical methods in diagnosis." Stat Methods Med Res 1992, 1:49-67.

Kreuz, M et al. "Development and implementation of an analysis tool for array-based comparative genomic hybridization." Methods Inf Med 2007, 46:608-613.

Larranaga, P et al.: Machine learning in bioinformatics. Brief Bioinform 2006, 7:86-112.

MacQuarrie, KL, Fong, AP, Morse, RH, and Tapscott, SJ. "Genome-wide transcription factor binding: beyond direct target regulation." Trends Genet 2011, 27:141-148.

Pelizzola, M et al. "MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment." Genome Res 2008, 18:1652-1659.

Sharp, AJ et al. "Methylation profiling in individuals with uniparental disomy identifies novel differentially methylated regions on chromosome 15." Genome Res 2010, 20:1271-1278.

Sharp, AJ et al. "DNA methylation profiles of human active and inactive X chromosomes." Genome Res 2011, 21:1592-1600.

Su, AI et al. "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A 2004, 101:6062-6067.

Toedling, J, Skylar, O, Krueger, T, Fischer, JJ, Sperling, S, and Huber, W. "Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts." BMC Bioinformatics 2007, 8:221.

Vapnik, VN. "An overview of statistical learning theory." IEEE Trans Neural Netw 1999, 10:988-999.

Wolpert, DH, and Wolf, DR. "Estimating functions of probability distributions from a finite set of samples." Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 1996, 54:6973.

Zhang, H, Liu, G, Chow, TW, and Liu W. "Textual and visual content-based anti-phishing: a Bayesian approach." IEEE Trans Neural Netw 2011, 22:1532-1546.