SciForum
Mol2Net

Genome-Wide Discriminatory Information Patterns of Cytosine DNA Methylation

Robersy Sanchez * and Sally A. Mackenzie *
N300 Beadle Center, University of Nebraska, Lincoln, NE 68588
* Author to whom correspondence should be addressed; E-Mails: robersy@unl.edu (R.S.); sally.mackenzie@unl.edu (S.A.M.)

Received: 5 October 2015 / Accepted: 5 October 2015 / Published: 2 December 2015

Abstract: Cytosine DNA methylation (CDM) is a highly abundant epigenetic heritable but reversible chemical modification to the genome. Herein, a machine learning approach, was applied to analyze the accumulation of epigenetic marks in 150 methylomes from *Arabidopsis thaliana* ecotypes. We hypothesize that these marks are chromosomal footprints that account for different ontogenetic and phylogenetic and histories of individual members of the sampling population. Our results support this hypothesis and suggest a statistical-physical relationship between CDM changes and single nucleotide polymorphism (SNPs). Furthermore, the genome-wide redistribution of CDM changes ensures the thermal stability of the DNA molecule preserving the integrity of the genetic message continuously stressed by thermal fluctuations in the cell environment.

Keywords: Epigenetics, Epigenomics, Information thermodynamics, Linear discriminant analysis

1. Introduction

Cytosine DNA methylation (CDM) is one of the molecular processes that result in epigenetic modifications to the genome. In particular, cytosine methylation is a widespread regulatory factor in living organisms. Changes introduced by DNA methylation can be inherited from one generation to the next. Some methylation changes can regulate gene expression and cause genomic imprinting [1,2]. Cytosine methylation arises from the addition of a methyl group to a cytosine's C5 carbon residue. Distinct pathways regulate methylation status by the action of methyltransferases [3]. The addition or removal of a methyl group to a cytosine C5 residue produces a change of information that is recognized by the molecular transcription machinery and can be verified by current sequencing technologies [2]. However, it is still undefined whether or not the observed methylation changes could be linked to genome-wide information patterns.

The development of DNA bisulfite conversion methodology coupled with next-generation sequencing approaches (Bis-seq) allows determination of the methylation status of nearly every cytosine in a genome. In this way, the methylation status of particular cytosine sites is

often expressed in terms of methylation level $p_i = \#C_i/(\#C_i + \#nonC_i)$, where $\#C_i$ and $\#nonC_i$ represent the numbers of methylated and non-methylated read counts observed at the genomic coordinate i, respectively. At tissue level, the methylation status (methylated or non-methylated) of cytosine C_i at the genomic coordinate i can be analyzed as a random variable that takes value "methylated" with probability p_i and "non-methylated" with probability $1-p_i$. Then, the formula $H(p(x_i)) = -\sum_i p(x_i) log_2 p(x_i)$ (1) of Shannon's entropy of a random event with probability distribution $p(x_i)$ can be applied to estimate the uncertainty of the methylation events at given cytosine site i as: $H(C_i) = -p(C_i) log_2 p(C_i) - (1-p(C_i)) log_2 (1-p(C_i))$ (2).

The entropy defined by Eq. 2 is therefore the expected value of the logarithm base 2 of the methylation level [4]. Assuming that, as a result of variations in environmental conditions, a change of methylation status in a genomic region *R* takes place, the uncertainty decrease in the genomic region *R* leads to a gain of information given by:

$$I_R = -\left(\sum_{i\in R} H\left(C_i^{after}\right) - \sum_{i\in R} H\left(C_i^{before}\right)\right) \quad (3)$$

Where C_i^{before} and C_i^{after} stand for the methylation status before and after the variations of environmental conditions, respectively [5]. Eq.3 expresses an information theoretical derived concept with a thermodynamic and biophysical meaning [5,6].

Herein, our study is focussed in the analysis of the genome-wide CDM information patterns induced by the changes in the enviromental conditions. In particular, we analyzed whether or not these information patterns carry discriminatory information in the form of chromosomal footprints.

2. Results and Discussion

The genome-wide evaluation of Eq.3 indicates the existence of methylation hotspots along the chromosomes (Fig.1). Genomic regions (GRs) can be classified according to the value of the I_R as: 1) highly variable methylation regions, 2) variable regions, and 3) low variable or constant regions. The regions with information gain (orange to black lines in the heatmaps color bar) or loss (light yellow to sky-blue) (Fig. 1) are observed at specific positions with a high line density in the pericentromeric region. The lines in yellow correspond to regions where the difference between the entropies $H_R^{ecotype}$ and H_R^{Col-0} is close to zero. According to Eq. 3, methylation hotspots are the ecotype chromosomal regions with a remarkable decrease in uncertainty with respect to Col-0.

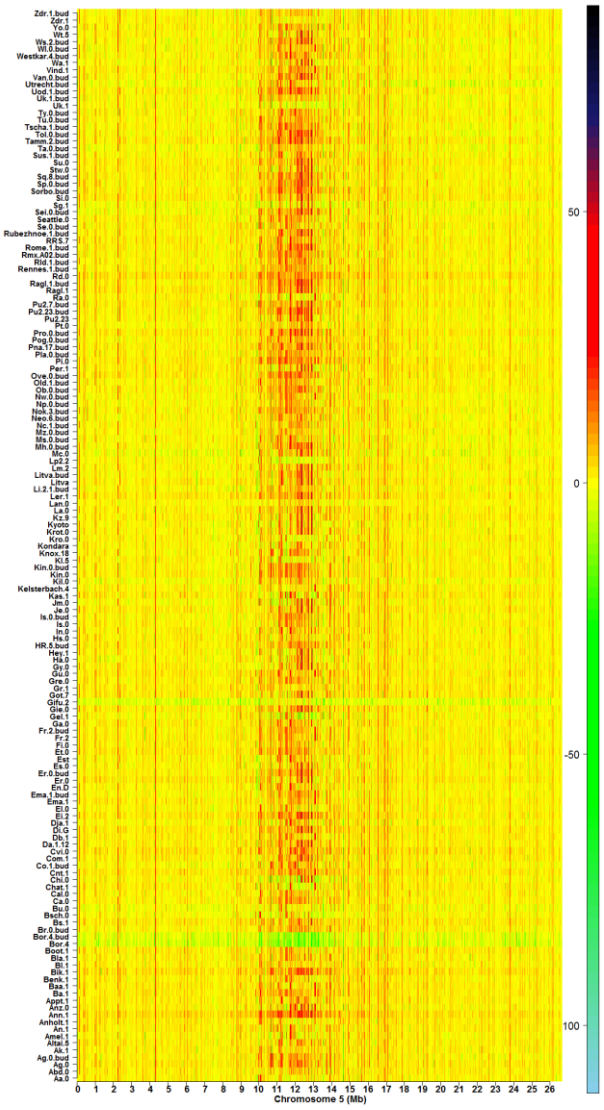
Methylation hotspots shared by a set of individuals at fixed chromosomal positions suggest the existence of specific informative landmarks (Fig. 1 and 2). That is, most of the CDM changes observed in natural variation and silencing mutants occur at specific methylation GRs, which are delineated in the heatmaps as chromosomal landmarks. These landmarks frequently cover transposable elements (TEs) and protein-coding regions (Fig. 2).

Discriminatory Informative Patterns in natural *Arabidopsis* ecotypes

The heatmaps suggest the existence of specific landmark informative patterns in all chromosomes across the ecotype samples that may or may not be shared by several individuals. These patterns comprise chromosomal regions carrying discriminatory information. That is, it is possible to distinguish between the individuals and among subsets of individuals by considering their discriminatory information patterns.

Figure 1. Methylation hotspots along chromosome 5 from 150 *Arabidopsis thaliana* ecotypes [7]. The color bar indicates the

magnitude of I_R values.



Applying hierarchical clustering based on the levels of C-DMRs, Schmitz *et al.* [7] found that the 150 *Arabidopsis thaliana* ecotypes from North America and Asia reflect their geographical distributions. Herein, the consecutive application of principal component analysis (PCA) and linear discriminant analysis (LDA) to the same ecotype set supports the hypothesis that the landmark patterns constitute chromosomal footprints that may account for ontogenetic and phylogenetic differences among individuals (Fig.3 A and C). This analysis supports not only the ecotype classification according to their geographical location for North America and Asia [7]), but also for all geographical regions.

These footprints are not only connected to the environment, but also to the single nucleotide polymorphisms (SNPs) detected throughout the DNA sequences (Fig.3 B and D). The classification of the *Arabidopsis thaliana* ecotypes according to their geographical distribution was retrieved not only from their landmark patterns, but also from their SNPs patterns (Fig.3). A summary of the classification results is presented in Table 1.

The similarity between the hierarchical clusters suggests that some statistical-physical relationship must exist between the SNPs and methylation changes. The two (2D) and three-dimensional (3D) kernel density plots presented in Fig.4 support the last hypothesis. The 2D kernel density plots indicate that the frequency of normalized read-counts supporting SNPs decrease with the increment of methylation changes, expressed here in terms of gain or loss of information I_R (Fig. 4A). This statistical trend is emphasized in the empirical 3D kernel density plots (Fig. 4B) and in the modelled Farlie-Gumbel-Morgenstern copula distribution built from the non-linear fit of the marginal distributions (Fig. 4C).

Figure 4 suggests that most of the observed CDM changes tend to preserve the integrity of the message carried by the DNA molecule, which is challenged by thermal fluctuations in the cell environment. This is consistent with the report that CDM changes alter the mechanical properties of the DNA molecule [8]. Thus, a statistical-physical relationship between CDM changes and SNPs is expected. Indeed, depending on the DNA sequence context, the addition or removal of a methyl group to a cytosine residue could increase or decrease the local thermodynamic stability of the DNA molecule and the nucleosomes [8–12]. The density plots of the experimental data indicate that the greatest frequency of SNPs is found in those GRs where the methylation status remains unchangeable with respect to the control (Col-0, Fig. 4).

Figure 2. Annotation of several hotspots on chromosome 2 from eigth *Arabidopsis* gene silencing mutants involving methylation.

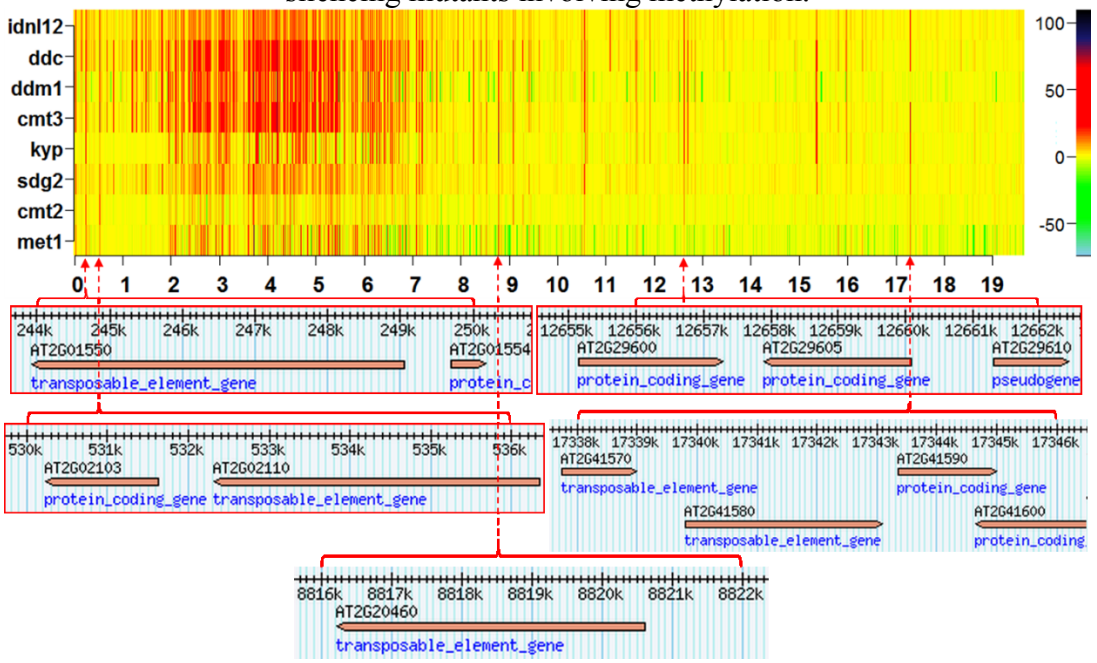


Figure 3. Classification of the Arabidopsis thaliana ecotypes according to their geographical distribution. **A** and **B**, LDAs based on I_R and SNPs, respectively. **C** and **D**, fan dendrograms based on the individual coordinates estimated from the LD functions. The dendrograms were built by applying hierarchical clustering with Euclidean distance and UPGMA as agglomeration method.

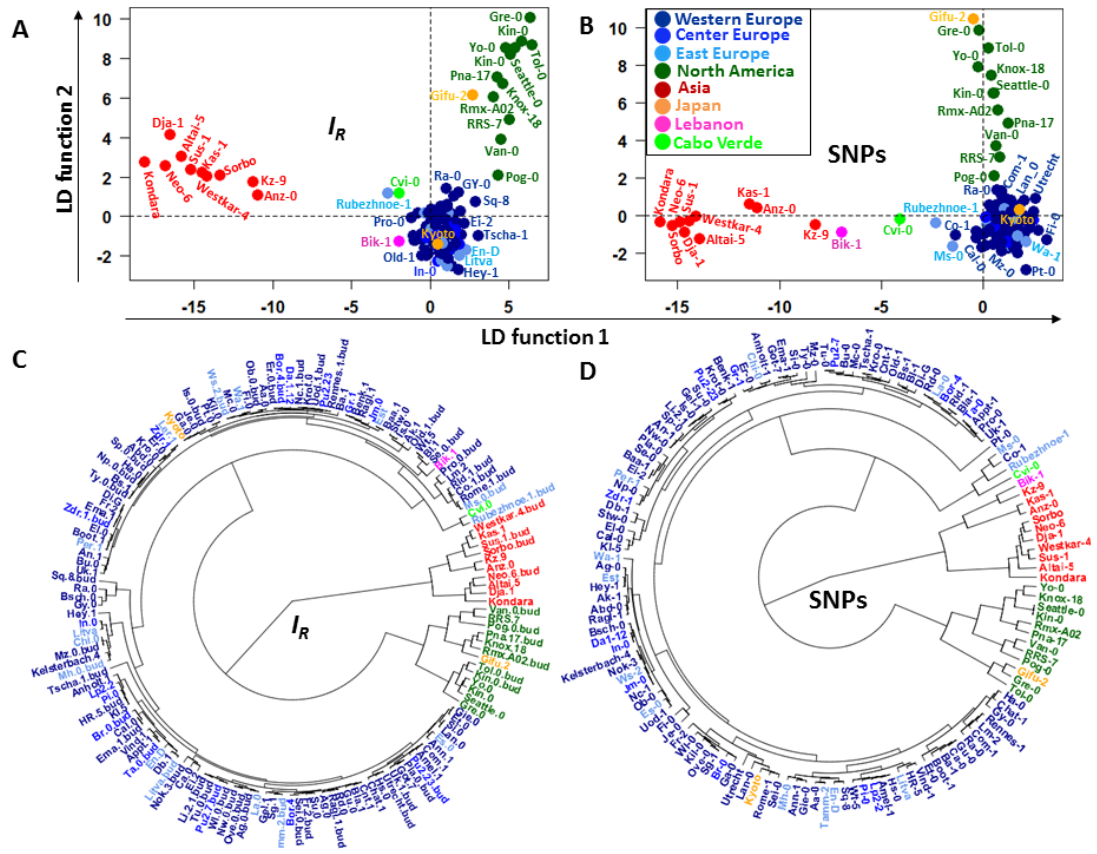
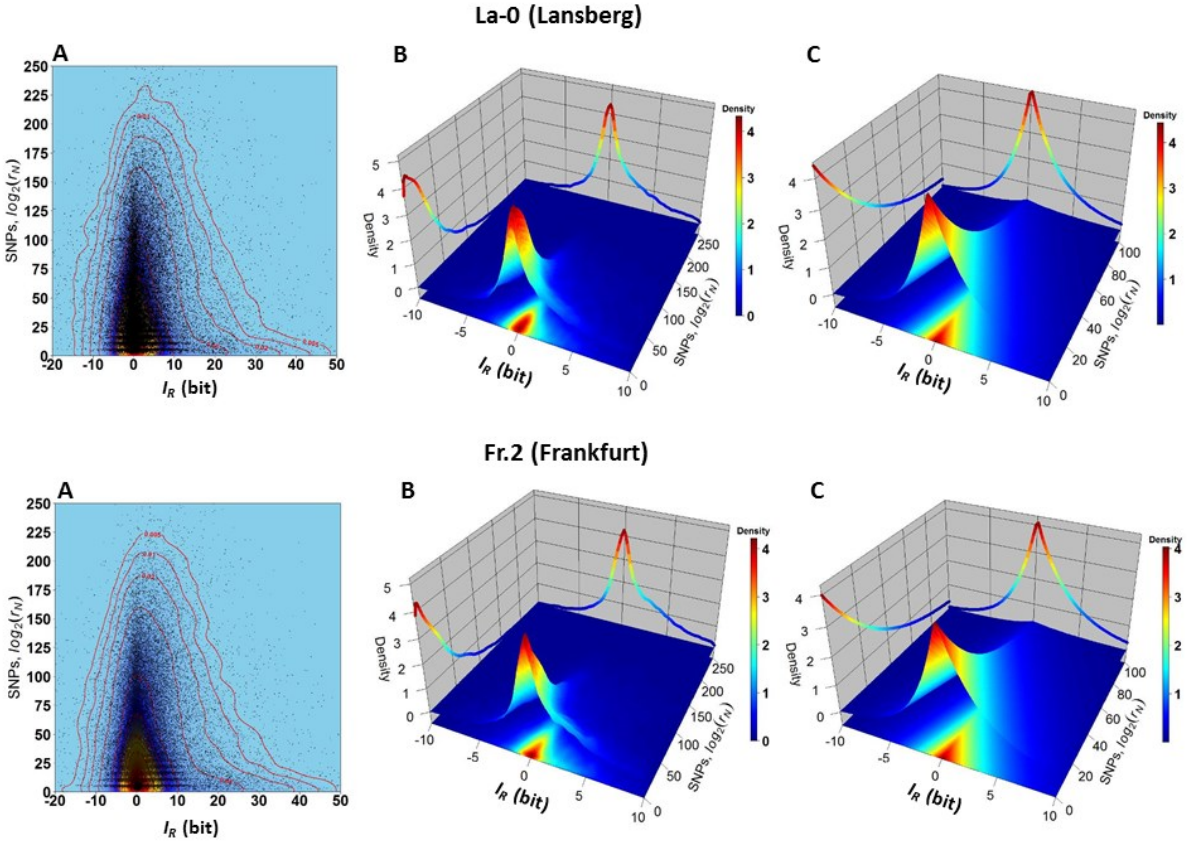


Table 1. Performance of the classifications presented in Fig. 3.

Sample [a]	Classifier	Accuracy Mean	2.5% quantile	97.5% quantile
CG ecotypes (2482 DIRs)	AUC+PCA+LDA	93.08352	88.05678	97.4359
	AUC+PCA+SVM	93.52517	91.83673	95.2381
	AUC+SVM	96.42381	95.91837	96.59864
SNPs ecotypes (2590 DIRs)	AUC+LDA	90.85758	85.42125	95.89744
	AUC+PCA+SVM	95.01642	94.02985	96.26866
	AUC+SVM	95.77007	95.23810	95.91837

[a] 1000 ten-fold cross-validations were performed for each classifier.

Figure 4. A: 2D kernel density plot. B: 3D kernel density plot. C: 3D plot of the density probability distribution of the Farlie-Gumbel-Morgenstern copula built from the non-linear fit of the marginal distributions estimated for LC_R (a Weibull PDF) and I_R (a Skew-Laplace PDF). These estimations were performed for several *Arabidopsis* ecotypes. The results for the ecotypes La-0 and Fr.2 are shown.



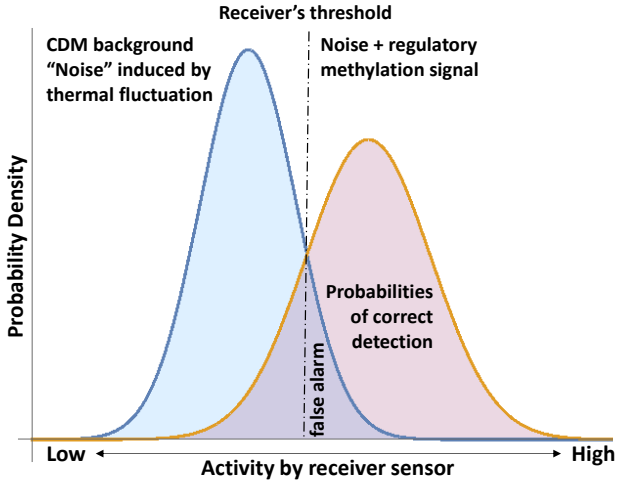
Hence, for an *Arabidopsis* plant, the adaptation to a new environment implies a genome-wide redistribution of CDM changes that will ensure the thermal stability of DNA. These are frequent methylation changes, which dynamically can vary from cell to cell in the same tissue. CDM changes

induced by thermal fluctuations are the simplest natural explanation to the "spontaneously occurring variations" of DNA methylation in *Arabidopsis thaliana* plants propagated by single-seed descent throughout generations [13,14].

An important subset of CDM changes regulates the process of gene expression and functional adaption to the environment [12]. These are specific molecular signals from the regulatory methylation machinery. At this point, the challenge is whether or not we would be able to sort out the regulatory methylation signals from the CDM background ("noise") induced by thermal fluctuations. This challenge has been already confronted (although in a different field, see references [15,16]) and a concrete application in the context of CDM is illustrated in Fig. 5. It is not possible to separate the regulatory methylation signal from the CDM background induced by thermal fluctuation. Even a simple regulatory methylation change could alter the mechanical properties of the DNA molecule [2,8,10] and, consequently, it could require an additional local readjustment. Therefore, the receiver (a device used by the experimenter to detect the signal) must set up a criterion for response, in this case, a threshold level of activity in its sensor (i.e., a function of the methylation levels). This threshold in combination with the PDFs for noise and signal plus noise determine the probabilities of correct detection [17] (Fig. 5).

Hence, any statistical analysis of the regulatory signals of CDM changes must consider the statistical thermodynamics subjacent to the methylation process. This concept conveys a suitable approach to discriminate the regulatory signals from the "noise" induced by the thermal fluctuations.

Figure 5. Signal detection in noise according to reference [15,16] and, here, applied to the detection of regulatory CDM signals.



3. Materials and Methods

Equation 3 was used to compute the I_R for several samples with methylation data available in online databases (see below).

Arabidopsis thaliana methylation data

According to Eq. 3, I_R is computed for a subject sample with respect to a given reference sample. The I_R values were computed for 150 Arabidopsis ecotypes [7]. The TSV files taken from NCBI GEO under accession GSE43857 [7] were read and transferred to R software version 3.2.1 [18] by using the Bioconductor (version 2.14) R-package *GenomicFeatures* [19]. Ecotype Col-0 was used as reference (152 ecotypes including Col-0). The mutant data used in Fig. 2 were reported in reference [20] (GEO accession numbers GSE39901).

Machine learning approach

To test the hypothesis that different environmental conditions must leave different landmark patterns on chromosomes, a machine learning approach was followed.

The estimation of the area under the ROC curve (AUC) for the current multiple-class classification problem was performed according to reference [21] and applied to reduce the space

dimension and to detected potential discriminant informative regions. This method was applied by using the R-package *HandTill2001*. Principal component analysis (PCA) was also used to reduce space dimensions.

AUC and PCA outputs were used with two classifiers: linear discriminant analysis (LDA) and support vector machine (SVM). These computations were performed by using the R-packages *adegenet* and *e1071*, respectively.

Logarithm of the normalized reads counts

The lists of SNPs and 1-bp deletions with a quality score of 25 and above of each ecotype samples were taken from 1001 Genomes Data Center (http://1001genomes.org/datacenter/; http://1001genomes.org/data/Salk/releases/). For a given number of non-repetitive reads supporting the base substitution (r), the normalized reads counts (r_N) were estimated as $r_N = r\ Concordance$, where $Concordance$ stand for the read ratios supporting a predicted feature to the total coverage. Next, the sum of logarithm base 2 of DNA-base substitution counts at a given region R was computed as:

$$LC_R = \sum_{i \in R} log_2\left(r_{N_i}\right) \quad 4.$$

4. Conclusions

The CDM changes observable at the heatmaps do not take place at random genomic regions, but at specific locations in chromosomes, hotspots of methylation changes, which are noticed as chromosomal landmarks in the heatmaps. The phylogenetic and ontogenetic history of each individual is reflected in the variations of landmark patterns, which like footprints carries discriminatory information about the individual.

Results indicate that, as a statistical tendency, most of the CDM changes preserve the thermodynamic stability of the DNA molecules. In addition, our study also leads to a new open practical problem: the discrimination between the regulatory methylation signals from the CDM background ("noise") induced by thermal fluctuations.

Acknowledgments

This work was supported by a grant from the Bill and Melinda Gates Foundation (OPP1088661) to S.M.

Conflicts of Interest

The authors declare no conflict of interest

References

1. Belanger AS, Tojcic J, Harvey M, Guillemette C (2010) Regulation of UGT1A1 and HNF1 transcription factor gene expression by DNA methylation in colon cancer cells. BMC Mol Biol 11: 9. doi:10.1186/1471-2199-11-9.

2. Dantas Machado AC, Zhou T, Rao S, Goel P, Rastogi C, et al. (2015) Evolving insights on how cytosine methylation affects protein-DNA binding. Brief Funct Genomics 14: 61–73. doi:10.1093/bfgp/elu040.
3. Law J a, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11: 204–220. doi:10.1038/nrg2719.
4. Shannon C. E (1948) A Mathematical Theory of Communication. Bell Syst Tech J 27: 379–423.
5. Schneider TD (1991) Theory of molecular machines. II. Energy dissipation from molecular machines. J Theor Biol 148: 125–137.
6. Tribus M, McIrvine EC (1971) Energy and Information. Sci Am 225: 179–188. doi:doi:10.1038/scientificamerican0971-179.
7. Schmitz RJ, Schultz MD, Urich M a, Nery JR, Pelizzola M, et al. (2013) Patterns of population epigenomic diversity. Nature 495: 193–198. doi:10.1038/nature11968.
8. Severin PMD, Zou X, Gaub HE, Schulten K (2011) Cytosine methylation alters DNA mechanical properties. Nucleic Acids Res 39: 8740–8751. doi:10.1093/nar/gkr578.
9. Římal V, Socha O, Štěpánek J, Štěpánková H (2015) Spectroscopic Study of Cytosine Methylation Effect on Thermodynamics of DNA Duplex Containing CpG Motif. J Spectrosc 2015: 1–8. doi:10.1155/2015/842810.
10. Yusufaly TI, Li Y, Olson WK (2013) 5-Methylation of cytosine in CG:CG base-pair steps: a physicochemical mechanism for the epigenetic control of DNA nanomechanics. J Phys Chem B 117: 16436–16442. doi:10.1021/jp409887t.
11. Portella G, Battistini F, Orozco M (2013) Understanding the connection between epigenetic DNA methylation and nucleosome positioning from computer simulations. PLoS Comput Biol 9: e1003354. doi:10.1371/journal.pcbi.1003354.
12. Flores KB, Wolschin F, Amdam G V (2013) The role of methylation of DNA in environmental adaptation. Integr Comp Biol 53: 359–372. doi:10.1093/icb/ict019.
13. Schmitz R, Schultz M, Lewsey M (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science (80-) 334: 369–373. doi:10.1126/science.1212959.
14. Becker C, Hagmann J, Müller J, Koenig D, Stegle O, et al. (2011) Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature 480: 245–249. doi:10.1038/nature10555.
15. Wiley RH (2006) Signal Detection and Animal Communication. Adv Study Behav 36: 217–247. doi:10.1016/S0065-3454(06)36005-6.
16. Wiley RH (2013) Signal Detection, Noise, and the Evolution of Communication. In: Brumm H, editor. Animal Communication and Noise. Berlin Heidelberg: Springer-Verlag, Vol. 2. pp. 7–31. doi:10.1007/978-3-642-41494-7.
17. Wiley RH (2013) A receiver–signaler equilibrium in the evolution of communication in noise. Behaviour 150: 1–37. doi:10.1163/1568539X-00003063.
18. R Core Team (2014) A language and environment for statistical computing.
19. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, et al. (2013) Software for computing and annotating genomic ranges. PLoS Comput Biol 9: e1003118. doi:10.1371/journal.pcbi.1003118.

20. Stroud H, Greenberg MVC, Feng S, Bernatavichute Y V, Jacobsen SE (2013) Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152: 352–364. doi:10.1016/j.cell.2012.10.054.
21. Hand DJ, Till RJ (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Mach Learn 45: 171–186.