BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 3. 419–430
Advance Access published on 6 November 2013
doi:10.1093/bib/bbt078

Applications of alignment-free methods in epigenomics

Luca Pinello, Giosuè Lo Bosco and Guo-Cheng Yuan

Submitted: 14th July 2013; Received (in revised form): 28th October 2013

Abstract

Epigenetic mechanisms play an important role in the regulation of cell type-specific gene activities, yet how epigenetic patterns are established and maintained remains poorly understood. Recent studies have supported a role of DNA sequences in recruitment of epigenetic regulators. Alignment-free methods have been applied to identify distinct sequence features that are associated with epigenetic patterns and to predict epigenomic profiles. Here, we review recent advances in such applications, including the methods to map DNA sequence to feature space, sequence comparison and prediction models. Computational studies using these methods have provided important insights into the epigenetic regulatory mechanisms.

Keywords: *epigenetics; nucleosome; DNA sequence; alignment-free method; machine learning*

INTRODUCTION

A fundamental biological question is to understand the function of the genome. Recent work has shown that the majority of the genome may be functional [1]. However, comprehensive functional annotation of the genome remains a daunting task. Development of computational methods has been essential for functional annotation. In particular, it has been a long tradition to use sequence alignment methods, such as BLAST [2] and FASTA [3], to generate initial biological hypotheses. In the past few years, the computational efficiency of sequence alignment has been improved greatly, enabling the rapid processing of large amounts of sequences [4,5].

Although the power of alignment-based methods has been well-demonstrated in a broad range of applications, their applications are not unlimited. When using a sequence-alignment method, one must make two basic assumptions:

(i) The functional elements share common sequence features; and

(ii) The relative order of these elements is conserved between different sequences.

However, the latter assumption is often violated for *cis*-regulatory elements, where there is little evidence suggesting the order between different elements would have any significant effect in regulating gene expression. In fact, many important regulatory elements are distal to transcription start sites, and it is the three-dimensional chromatin structure that facilitates their regulatory interactions [6]. As such, the recently developed alignment-free methods [7] have emerged as a promising approach to investigate the regulatory genome.

During the past decade, epigenetics has become increasingly recognized as an important layer of control of gene activities. Epigenetic regulation refers to heritable changes of gene expression that occur without alteration of the DNA base sequence [8]. Using an analogy, we can think that the genomic DNA represents all the potential functions of the 'software of the cell', while the epigenetic

Corresponding authors. Giosuè Lo Bosco, Dipartimento di Matematica e Informatica, Via Archirafi 34, 90123 Palermo-Italy. E-mail: giosue.lobosco@unipa.it; Guo-Cheng Yuan, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, MA 02215, USA. E-mail: gcyuan@jimmy.harvard.edu

Luca Pinello is Research Fellow at the Department of Biostatistics and Computational Biology at Dana-Farber Cancer Institute and at the Harvard School of Public Health, Boston, USA. His research is focused on new computational approaches to investigate different epigenetic factors that regulate the chromatin structure and dynamics.

Giosuè Lo Bosco is Assistant Professor of Computer Science at the Department of Mathematics and Computer Science at the University of Palermo, ITALY. His research focuses on the design and development of machine-learning tools.

Guo-Cheng Yuan is Associate Professor at the Department of Biostatistics and Computational Biology at Dana-Farber Cancer Institute and at the Harvard School of Public Health, Boston, USA. His research focuses on epigenomics and gene regulatory networks.

mechanisms dictate some control on which features are enabled/disabled in each particular cell type, specializing this 'software'.

While epigenetics and genetics were studied in isolation in early studies, recent studies have shown that they are indeed closely related [9]. It has been increasingly recognized that the genomic sequence plays an important role in guiding epigenetic pattern specificity. However, a major challenge is that, unlike transcription factors (TFs), most epigenetic factors do not or weakly interact with the DNA. While epigenetic factors can be brought to specific DNA sequences through interaction with TFs, they rarely have unique interacting factors [10]. As a result, their association with DNA sequence patterns is, in general, much weaker than TFs.

An emerging field of research is to apply alignment-free methods to study the link between the genome and epigenome. In particular, a number of alignment-free methods have been previously used to predict genome-wide epigenetic patterns and to identify the sequence elements that play a regulatory role in establishing epigenetic patterns. In this article, we review recent progress in this exciting direction.

EPIGENETICS PATTERN AND DNA SEQUENCES

Epigenetic mechanisms include nucleosome positioning, histone modifications and DNA methylation. The primary repeating unit of chromatin is the nucleosome, which consists of 147 bp of DNA wrapped 1.67 times around an octamer of core histone proteins [11]. The N-terminal ends of the histones are unstructured and called the histone tails. Many amino acid residues on the histone tails can be covalently modified, and many of these modifications have distinct biological functions [12]. An important task is to characterize the biological function of the combinatorial states of multiple histone marks [12]. Covalent modification can also occur at the DNA level [13], where the cytosine nucleotide can be methylated at the 5′ position.

The field of epigenomics has been growing rapidly in the past few years, in part because of the development of genome-wide profiling technologies [14–19] and new specialized computational approaches [20–22]. Previous studies have identified strong relationship between epigenetic patterns, TF binding and downstream gene activities. Dynamic changes of epigenetic patterns are strongly associated with developmental control, environmental response and disease susceptibilities. Several computational approaches have been developed to characterize the genome-wide chromatin states [23–28]. Insights obtained from these studies have been essential for developing new therapeutic approaches [29].

The underlying mechanisms for the establishment and maintenance of cell type-specific epigenetic patterns are complex and involve the dynamic interactions among multiple classes of factors, including chromatin modifiers, DNA binding proteins, non-coding RNAs and signaling molecules [10,30]. However, the relative contribution of each factor remains poorly understood. One fundamental question is to what extent the epigenomic patterns are orchestrated by the underlying DNA sequence. Previous studies have shown that, at least for a subset of epigenetic marks, the DNA sequence plays an important role in targeted recruitment to specific genomic regions [31]. Distinct DNA sequence features have been identified to be associated with nucleosome positioning [32], DNA methylation [33] and histone modifications [34], and methods have been developed to predict genome-wide patterns, sometimes with great accuracy. Such a strong association with DNA sequence features may be surprising at first sight, but it is not incompatible with the fundamental property that epigenetic patterns may change without change of DNA sequences. What this means is that the DNA sequence may mark specific regions where epigenetic pattern changes are likely to occur, but the cell type-specific patterns still depend on the activities of chromatin regulators and mediating TFs, which may be further regulated by inputs from various signaling pathways or other mechanisms. Interestingly, the most informative sequence features are those that are traditionally viewed as 'degenerative', such as CpG density and poly-A tract [9], posing a severe challenge for alignment-based methods. In the meantime, the development of alignment-free methods has provided a promising alternative to overcoming such challenges.

MAPPING SEQUENCE TO FEATURE SPACE

To identify common sequence features associated with an epigenetic pattern, a major task is to compare multiple DNA sequences. Traditionally,

sequence comparison is achieved by multiple sequence alignment and quantified by a quality score [35]. However, it is infeasible to align a large set of sequences because of overwhelming computational cost. Furthermore, for sequence-alignment methods to be effective, one must assume that the underlying sequences are highly conserved, which is often violated in epigenomics.

In contrast, alignment-free methods make no assumption regarding sequence conservation or the order between elements [7]. To compare different sequences, a first step is to project each sequence into a suitably selected feature space, where sequence information is transformed to numerical information, thereby enabling the application of various established mathematical and probabilistic tools for comparison.

In particular, a DNA sequence s is represented as a string of L symbols from a finite alphabet Σ of length r. We indicate as Σ^* the set of all the strings that can be obtained by finite successions of symbols in the alphabet Σ. Any generic string s is mapped to a vector $\boldsymbol{x}_s = \varphi(s)$ (the feature vector) by a particular mapping function $\varphi : \sum^* \rightarrow F$, allowing a representation of s into a multidimensional space (the feature space) F, where a particular distance function between the vectors can be adopted to reflect the observed similarities between sequences. Here we summarize a few commonly used representations of feature spaces.

Mapping by word composition and *k*-mers

One of the most common ways of defining sequence features is by enumerating the frequency of occurrence of a finite set of preselected words. The simplest and most common choice of word list is k-mers, i.e. any string of length k whose symbols are taken in the oligonucleotide alphabet (A,T,C,G). In this case, each sequence s is mapped to a vector in a 4^k dimensional feature space.

In general, each numerical component x_s^i of the feature vector is set to the value f_i^s that represents the frequency of the i-th k-mer w_i in s, each one counted by a sliding window of length k that is run through the sequence s, from position *1* to $L-k+1$. Another possible choice is to set x_s^i to the empirical probability $p_i^s = f_i^s/(L-k+1)$, leading to the application of information-based distances.

The mapping of sequences to the space of k-mers is widely used in epigenetic context; it has been used to study the nucleosome positioning [36–38], to characterize the complexity of the sequence associated with particular histone modifications [39], to predict histone modification patterns [40] and to predict DNA methylation patterns [41–43].

Mapping to suffix tree

Because of the large number of possible k-mers, it is often desirable to reduce the complexity by selecting a subset of k-mers for detailed analysis. One of the commonly used techniques to achieve this goal is to use efficient data structures for string operations like suffix tree [44], Burrows–Wheeler transform [45], suffix array [46] and compressed index [47].

A suffix tree T of a string s on a finite alphabet $\sum$ is defined as a tree containing all the suffixes of s. In particular, each suffix of s corresponds to one path in T. Given a generic string s of length L_s, its suffix tree can be constructed in linear time $O(L_s)$ and the search of any string t with length $L_t \leq L_s$ in s is $O(L_t)$ (this result holds also for a finite number of mismatches in t). For these reasons suffix trees and related data structures are widely used in sequence-alignment software, such as BWA [48] and BOWTIE [49]. In contrast, in the alignment-free context these structures are used to efficiently evaluate distances or similarity measures based on word composition, where the computational cost scales linearly with respect to the word length. Of particular interest is a distance function proposed by Apostolico *et al.* [50] based on the enumeration of all the possible k-mers of different and finite length in linear time. A similar idea is used by another study, which defines a general similarity measure [51]. In particular, this measure does not compute all the possible k-mers present in the sequences, but instead calculates only the optimal word length to use. In addition, suffix trees have been used to define distance functions based on compressibility by extensible motifs, i.e. motifs that may contain a finite number of gaps [52].

Suffix trees have been used in the context of epigenomics, for example, to predict hypomethylated regions and CpG islands using hexamer alphabets [53] and to investigate the DNA sequence periodicity associated with nucleosome positioning [54]. In the latter study, Rasheed *et al.* [55]applied suffix trees to detect repeating dinucleotide subsequences in nucleosomal DNA, using as input a list of known chicken and yeast nucleosomal DNA sequences in the literature. Random sequences extracted from

the respective genomes are used as control. Periodicity is detected by searching for subsequences containing recurrent dinucleotides at the following positions:

$$i, i+p \pm W, i+2*p \pm W, \ldots, i+j*p \pm W$$

where p is the period, i a position in the sequence, j an integer and W is the width of the sliding window. They found that around 90% of the nucleosome sequences contain periodic subsequences accordingly to the above criterion, while only 10% of the random sequences have such a property. In this application, the application of suffix tree is essential for achieving sufficient numerical efficiency.

Mapping into spatially extended signal space

A sequence s of length L can also be mapped into a spatially extended signal space. Starting from a finite set of n fixed word length segments w_i, the component x_s^i is a one-dimensional binary digital signal defined by:

$$x_s^i(t) = \begin{cases} 1 & s_t s_{t+1} \ldots s_{t+k} = w_i \\ 0 & otherwise \end{cases}$$

where k is the length of w_i.

An advantage of using this kind of representation is that it takes into account the spatial 'shape', which may have favorable binding energy. There are a large number of digital signal processing techniques like convolution, classical Fourier or wavelet analysis that can be used to discover spatially extended patterns.

In the particular problem of nucleosome positioning prediction, the signal representation has shown particular merits. Starting from a training set of nucleosome sequences, the probability density P_{w_i} of a dinucleotide w_i at each position of a sequence underlying a nucleosome is estimated. Finally, a convolution operation between P_{w_j} and x_s^j is used to score if s underlies a nucleosome [56].

A wavelet transform is a decomposition of a spatially extended signal into the space–frequency domain via a set of basis functions, each related to a common function, called the mother wavelet, by scaling and translation. Wavelet transforms can be viewed as an extension of traditional Fourier analysis in that the signal of interest does not need to be periodic. Yet, similar to Fourier analysis, wavelet methods can be used to systematically investigate the shape of a signal in terms of different frequencies. Wavelet methods have been widely used in many areas of science and engineering [57].

To study the role of DNA sequence periodicity in determining nucleosome positioning, Yuan and Liu [37] applied wavelet analysis to predict nucleosome positioning in yeast by analyzing the DNA sequence information from 199 nucleosomal sequences and 296 linker sequences known in the literature [37,55]. They further truncated each input nucleosome or linker DNA sequence s to 129 bp long. Dinucleotide frequencies, w_i are transformed to wavelet coefficients x_s^i for $i = 1, \ldots, 16$ using the Haar wavelet basis, defined as

$$\Psi(t) = \begin{cases} 1 & for\ 0 < t < 1/2 \\ -1 & for\ 1/2 < t < 1 \\ 0 & otherwise \end{cases}$$

This transform involves seven layers of wavelets, each layer corresponding to a specific length scale. The sum of square of the wavelet coefficients at each layer, also called the wavelet energy, is then used to summarize the sequence features with the aim to quantify the contribution of different spatial frequencies. They find modest prediction power (AUC = 0.84) in predicting nucleosome positioning, but the nucleosome-depleted regions are highly predictable.

A major benefit of the wavelet approach is its ability to detect position-specific signals across multiple length scales, allowing combinations of both small-scale and large-scale motifs to contribute to the overall targeting signal. To date, the wavelet approach has only been applied to investigate the spatial distribution of dinucleotides. In general this can be naturally extended to longer k-mers, but there is an additional computational cost compared with simply using word counts.

COMPARISON OF SEQUENCES IN THE FEATURE SPACE

Once DNA sequences are mapped to numerical vectors in a feature space F, they can be easily compared by using traditional mathematical tools, such as a distance or dissimilarity function, d, which is a two-variable function on F satisfying the following conditions:

(1) $d(\boldsymbol{x},\boldsymbol{y}) \geq 0$
(2) $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x})$
(3) $d(\boldsymbol{x},\boldsymbol{x}) = 0$

Generally speaking, there are three main classes of distance functions that have been applied to epigenetic studies: geometrical, correlation and information-based distances.

Geometrical distance functions capture the concept of physical distance between two objects. They are strongly influenced by the magnitude of changes in the measured components of vectors x and y, making them sensitive to noise and outliers. The most adopted geometrical distance is the 'Euclidean distance':

$$d_e(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^t}$$

Correlation distance functions capture linear dependencies between the coordinates of two vectors. The most adopted distance belonging to this class is the 'Pearson's correlation distance':

$$d_\rho(\boldsymbol{x},\boldsymbol{y}) = 1 - \left|\rho_{\boldsymbol{x},\boldsymbol{y}}\right|$$

where $\rho_{x,y}$ is the Pearson correlation between x and y.

Information-based distance functions are defined via well-known quantities in information theory, such as entropy or relative entropy. They have the advantage of capturing nonlinear statistical dependencies between two discrete variables. The entropy measures the degree of uncertainty associated with a random variable and typically is represented in units of bits. The mathematical definition is as follows.

Given a multivariate random variable X with possible values $\{x_1,\ldots,x_n\}$ and its associated probability mass function $P(X)$, the entropy is defined as

$$H(X) = -\sum_{i=1}^{n} p_i^X \log_2\left(p_i^X\right)$$

where $p_i^x = P(x_i)$.

The concept of entropy can be extended to 'relative entropy' or 'Kullback–Leibler divergence', which quantifies the similarity between two random variables X and Y:

$$RH(X,Y) = \sum_{i=1}^{n} p_i^X \log_2\left(\frac{p_i^X}{p_i^Y}\right)$$

While relative entropy is not symmetric, it can be symmetrized by the following transformation:

$$d_{kl}(X,Y) = \frac{RH(X,Y) + RH(Y,X)}{2}$$

and the result is called the Kullback–Leibler distance. To evaluate the Kullback–Leibler distance between a pair of DNA sequences, the occurrence of each k-mer is viewed as a random variable following a binomial distribution whose parameter is estimated by its overall frequency.

The idea described above has been used to define an unbiased quantitative measure for DNA sequence specificity called the Motif Independent Measure (MIM) [39]. Starting from a set of random sequences, the method uses their empirical probabilities of k-mers (default $k=4$) to estimate their Kullback–Leibler divergence distribution, and uses this distribution as the null model to establish the level of specificity. Of note, this approach does not assume the existence of any distinct sequence motif. The authors apply this approach to analyze a published ChIP-seq dataset of the enhancer mark H3K4me1 [1,58,59]. They find that the sequence specificity associated with the H3K4me1 peaks is highly cell type-specific, which is highest in embryonic stem cells, suggesting different targeting mechanisms. Of note, no distinct motif sequences are identified in the H3K4me1 targets, suggesting a motif-independent utility of this approach in evaluating sequence specificity.

The concept of entropy has also been applied to investigate the relationship between nucleosome positioning and DNA sequences. For example, Levitsky *et al.* [60] identify differences in entropy levels between different classes of nucleosomal DNA sequences, which may be associated with conformational and physicochemical property differences, based on analysis of an existing nucleosome sequence database [61].

Kernels for sequences comparison

Another class of sequence comparison methods is the string kernels, which have recently gained significant popularity and been successfully applied to a wide range of problems [62,63].

Consider two generic sequences s and t, and a generic feature space F where an inner product is defined, a kernel is any $K : \sum^* \times \sum^* \to \Re$ such that there exists a mapping $\varphi : \sum^* \to F$ satisfying:

$$K(s,t) = \langle \varphi(s), \varphi(t) \rangle$$

where $\langle \cdot,\cdot \rangle$ denotes the inner product in F.

Classical examples of kernels are:

$K(x,y) = x^t y + c$	*Linear*
$K(x,y) = (\alpha x^t y + c)^d$	*Polynomial*
$K(x,y) = e^{-(\gamma \|x-y\|)}$	*Gaussian*

It is straightforward to transform a kernel to a distance function, as follows:

$$d_K(s,t) = \sqrt{K(s,s) + K(t,t) - 2K(s,t)}$$

In computational biology, a popular choice is subsequence kernel, which is defined in a number of steps. First, given a sequence $s = s_1 s_2 \ldots s_L$, the *index set* $\boldsymbol{i} = i_1 i_2 \ldots i_l$ of length l is defined as a monotonically increasing set of positive integers whose values are no greater than L. Denote by $s[\boldsymbol{i}] = s_{i_1} s_{i_2} \ldots s_{i_l}$ the subsequence (allowing gaps) of s that corresponds to the index set. Second, each index set $\boldsymbol{i}$ is associated with a gap-penalizing weight, $\lambda^{l(\boldsymbol{i})}$, where $l(\boldsymbol{i}) = i_l - i_1 + 1$ and $\lambda < 1$ is a predetermined constant. Third, the sequence s is mapped into a $|\Sigma|^m$ dimensional feature space through the following mapping:

$$\varphi_u(s) = \sum_{\boldsymbol{i}:s[\boldsymbol{i}]=u} \lambda^{l(\boldsymbol{i})}$$

for every subsequence u of length m. Finally, the subsequence kernel is defined as the dot product between any pair of sequences:

$$K_m(s,t) = \sum_{u \in |\Sigma|^m} \sum_{\boldsymbol{i}:s[\boldsymbol{i}]=u} \sum_{\boldsymbol{j}:t[\boldsymbol{j}]=u} \lambda^{l(\boldsymbol{i})+l(\boldsymbol{j})}$$

As such, indels are tolerated yet penalized in a suitable manner.

Because of the poor numerical efficiency, effort has been taken to develop a more efficient algorithm based on recursive strategies. As a result, the numerical complexity is improved to $O(mL_sL_t)$ [64], where L_s and L_t are the lengths of the two considered sequences.

Sequence comparison using complexity/compression measures

Another class of distance functions has been recently introduced based on an extension of the Kolmogorov complexity, which intuitively can be viewed as a measure of the computational resources needed to generate such sequences. The mathematical definition is as follows.

The conditional Kolmogorov complexity $KC(s|t)$ between two generic strings s and t is the length of the shortest binary program written in a generic programming language that computes s giving t as input. The Kolmogorov complexity $KC(s)$ of a string s is defined as $KC(s|\lambda)$ where λ stands for the empty string.

Li *et al.* [65] define the universal similarity metric (USM) for strings based on the Kolmogorov complexity:

$$USM(x,y) = \frac{max\{KC(s|t^*), KC(t|s^*)\}}{max\{KC(s), KC(t)\}}$$

where w^* denotes the shortest program that produces the generic sequence w on an empty input. USM represents a lower bound for all the distance and similarity functions.

It is difficult to exactly compute the value of USM; therefore, it is commonly approximated by the following formula:

$$UCD(s,t) = \frac{max\{|C(st)| - |C(s)|, |C(ts)| - |C(t)|\}}{max\{|C(s)|, |C(t)|\}}$$

where st and ts denote the concatenations of the sequences s and t, C is a compression algorithm and $C(s)$ its output on a string s. Details about this approximation can be found in the work of Ferragina *et al.* [66]. Complexity-based approaches have been used, for example, to distinguish nucleosome-enriched and -depleted regions [67].

PREDICTION OF EPIGENETIC PATTERNS IN SEQUENCE FEATURE SPACE

While epigenetic patterns are typically not strongly associated with any distinct DNA sequence features, accurate prediction models have been built by combining multiple sequence features. There are three main classes of approaches, as described below.

Regressors and classifiers

Regression analysis is the process of determining how a variable y (the response) is related to a multivariate variable $\boldsymbol{X}$ (whose components are called covariates). Assuming that y and $\boldsymbol{X}$ are associated via a prespecified, possibly nonlinear functional relationship $y = h(\boldsymbol{X}, \beta)$ with an unknown coefficient β, the regression procedure infers the values of β based on a finite number of observed data. Classification is similar to regression but for discrete values of y. The choice of a particular function form of h is related to the so called empirical risk:

$$R_e = \frac{1}{m} \sum_{i=1}^{m} L\big(h(X_i), y_i\big)$$

where $L(x,y)$ is the loss function such that it is non-negative and $L(x,x) = 0$. Empirical risk

minimization is the process of choosing a function $\hat{h}$ that minimizes the empirical risk:

$$\hat{h} = argmin_{h \in H}\ R_e\ (h)$$

where H is a particular class of function. In binary classifications, the linear hyperplanes $\beta^t X$ can be used to define the set H of binary linear classifier as:

$$H = \left\{sgn\left(\beta^t X\right) : \beta \in \Re^n\right\}$$

A commonly used regressor is the logistic regression (LR), which is a member of the generalized linear models suitable for binary responses [68]. In LR one obtains the logarithm of the odds of the success outcome, and finally uses a log-linear model:

$$ln\left(\frac{p}{1-p}\right) = \sum_j \beta_j x_j$$

where p represents the probability of the success outcome, x_j represents the covariate values, and the β_js are unknown coefficients that are numerically estimated, often with the maximum likelihood procedure.

In the context of epigenomics, Yuan and Liu [37] use wavelet energies as covariates of an LR model to predict the likelihood of an arbitrary sequence being classified as a nucleosomal or a linker sequence. They find that the genome-wide nucleosome-depleted regions can be predicted with high accuracy from the DNA sequence alone.

In Yang *et al*. [41] an LR model, based on *k*-mers, is used to study cell type-specific DNA methylation susceptibility at three different resolutions: promoters, CpG dinucleotides and CpG segments. They predict, with reasonable accuracy, methylation patterns across different human cell types. Furthermore, they also identify *k*-mers that partially explain the tissue-specific methylation patterns.

Support vector machines

Support vector machines (SVMs) [69] are binary linear classifiers. They search for the optimal hyperplane $\hat{h}$ that maximizes the geometric margin, which is the distance of the hyperplane to the nearest training data point of any class. The main advantage of SVM is that it provides a solution to the global optimization problem, thereby reducing the generalization error of the classifier.

The formulation of SVM can be easily extended to build a nonlinear classifier by incorporating a generic kernel of the class H:

$$H = \left\{sgn(K(\beta, X)) : \beta \in \Re^n\right\}$$

While no systematic tools have been developed to automatically identify the optimal kernel for a particular application, intuition-based kernel designs often seem to be sufficient in practice.

Peckham *et al*. [36] have applied SVM to predict nucleosome positions in yeast from DNA sequences using genome-scale nucleosome positioning information identified by a tiling array experiment [20]. In particular, each input sequence is represented by *k*-mers with lengths up to 6 bp, using a linear kernel. Genome-wide nucleosome positions are predicted using a hidden Markov model to detect peaks of the SVM-predicted scores. Similarly, using a linear kernel, SVM has also been applied to predict histone modifications patterns [40].

In addition, Lee *et al*. [70] have applied SVM to predict enhancer locations. In this study, they used ChIP-seq data of EP300/CREBBP [71] to define enhancers, and also used a linear kernel for prediction. Interestingly, a subset of the identified sequence elements can be associated to sequence motifs of TFs that are known to play a tissue or developmental stage-specific role in gene regulation.

Classification trees and random forests

Classification trees are complex classifiers based on the idea of representing the function h as a set of discrimination rules in a feature space in term of discrete data structures called rooted trees [72]. A rooted tree is a direct and acyclic graph composed of a set of nodes and a set of paths that connect the nodes. The root is a special node where all paths ultimately originate from, whereas the leaf nodes are where the paths terminate. Each node represents a binary rule that splits the feature space according to the value of a predictive feature, and a path from the root to leaf nodes represents a series of rules that are used to recursively divide the feature space into smaller subspaces, where a common class label is assigned. This approach can also be extended using an ensemble of trees, called forests, to make classification, and the results from individual trees are aggregated, for example, by majority voting. The most used methods in this context are decision trees [73], random forests [74] and probabilistic sum of trees [75]. The tree-based models have been used successfully, for example, to predict target regions of a chromatin remodeling complex based on TF binding and motif sequences [76] or for the identification of DNA binding protein [77].

An ensemble version of classification trees, called the Bayesian additive regression trees (BART), has

been developed by Chipman *et al.* [75]. Previous work has shown that the BART model has competitive performance with other methods, such as SVMs [78]. An advantage of BART is that, by enumerating the frequency of usage of each predicting variable, it can serve as an effective way for variable selection.

Liu *et al.* [76] apply the BART model to predict the genome-wide polycomb target genes in mouse embryonic stem cells. The Polycomb group complexes (PRC1 and PRC2) are responsible for synthesizing the repressive H3K27me3 mark and play an important role in development. ChIP-chip experiments identify nearly 3000 genes whose promoters are occupied by PRC2 complex in mouse embryonic stem cells; these target genes are highly enriched for developmental regulators, which are activated during cell differentiation in a cell type-specific manner. Using the BART model, Liu *et al.* [76] compare the promoter sequences of the target and nontarget genes and find that the model has good classification accuracy (AUC = 0.83).

There are a number of variations of tree-based models in the literature using different learning strategies and, as a result, the computational complexities are very different. For example, classic approaches like the ID3 or C4.5 [79] have a complexity of $O(n^2\ m)$ in the case of n discrete features or $O(n^2\ m^3)$ for n continuous features [80]. Random forests instead have a complexity of $O(tnmlogm)$, where t is the number of trees in the forest. The complexity of the BART approach instead is dominated by the Markov chain Monte Carlo sampling, which has been shown to have, in the case of large samples, a stochastic complexity of $O(n^2)$ [81]) and is $O(t\,m\,n^2)$.

In contrast, the complexities of SVM and LR can be highly variable depending on the specific optimization problem. Fast implementation of the LR optimizer involves a computation that is $O(m\,n^2)$ [82].

In the case of SVM, the optimization problem is a quadratic-programming one, and it is noteworthy that the best solvers have $O(m^3)$ complexity [83], in addition to the cost, $k(n)$, associated with the precomputation of the kernel matrix. For the linear kernel, $k(n)$ is equal to n^2, but this cost can be significantly greater for more complex kernels. Numerically efficient implementations have been developed [83].

To date, effective prediction models have been developed for nucleosome positioning, histone modification and DNA methylation, providing strong evidence that DNA sequence plays an important role in guiding these factors. A summary of the methods discussed in this article is presented in Table 1.

As summarized in Table 2, each method has its strengths and weaknesses. For example, SVM models provide usually highly competitive prediction accuracies, but the results are often difficult to interpret in terms of the underlying features. On the other hand, the tree-based models may perform less efficiently in terms of prediction accuracy, but their results can be easily interpreted. Furthermore, some methods, such as LR, are more numerically effective than others, although not parallelizable. These properties should be taken into consideration for model selection.

Table 1: A summary of alignment-free methods that either have been applied to epigenomics or may be suitable for epigenomics

Epigenetic mark	Feature space mapping	Similarity/distance function	Prediction model
Nucleosome positioning/occupancy	Word composition [36,37,38]	Kernel [36]	Logistic regression [37]
	Suffix tree [54]	Information based	Support vector machine [36]
	Signal space [37,56]	Complexity based [67]	Tree models
	Entropy [37,60]		
DNA methylation	Word composition [41,42,43]	Kernel [43]	LR [41,43]
	Suffix tree [53]	Information based	Support vector machine [43]
	Signal space	Complexity based	Tree models
	Entropy		
Histone modifications/ chromatin regulatory elements (enhancers)	Word composition [39,40,70,76,34]	Kernel [40,70]	LR [34]
	Suffix tree	Information based	Support vector machine [70,40]
	Signal space	Complexity based	Tree models [76]
	Entropy		

Table 2: A comparison of prediction models that have been applied to epigenomics

Prediction model	Time complexity	Parallelizable	Easy to interpret
SVM/kernel based	$O(m^3)$ for the QP solver and $O(k(n))$ for the computation of the kernel matrix, where $k(n)$ is the computational cost of the kernel used	Yes	No
LR	$O(m\,n^2)$	No	Yes: it is easy to check the contribution of each feature in the model.
Tree based	• ID3 or C4.5: $O(n^2\,m)$ discrete features, $O(n^2\,m^3)$ continuos features • BART: $O(t\,m\,n^2)$ • Random Forest: $O(tn\,m\,log\,m)$	Yes	Yes: it is easy to check the contribution of each feature and also combinatiorial rules in the features space.

DISCUSSION

In this article, we have reviewed a number of recent applications of the alignment-free methods to epigenomics, whose connection with the DNA sequences has not been well recognized until recently. Because of the limit of space and our knowledge, we have only touched upon a small set of existing studies and apologize for the omission of any important studies. Nonetheless, it is evident that the alignment-free methods have offered new biological insights into DNA methylation patterns [41–43], histone modifications [70] and nucleosome positioning [37,38]. We think that the power of alignment-free methods mainly lies on their flexibility, robustness and numerical efficiency.

Most previous studies either focus on a specific epigenetic mark or consider different marks independently, but in reality there is often strong association between different marks. For example, it is well known that nucleosome positioning influences DNA methylation patterns [84]. Similarly, DNA methylation is known to be positively correlated with H3K9 methylations and negatively correlated with H3K4 and H3K27 methylations [17,85]. Taking these correlations into account may result in more powerful predictive models.

The DNA sequence is only one of many factors that affect epigenetic patterns. The effect of additional factors, such as transcription, replication and histone-mediated self-propagation, varies greatly among different epigenetic marks, providing an intrinsic limit for any sequence-based prediction model. One important question is whether the lack of prediction power is because of the inferiority of computational method or such intrinsic limitations. This is in part addressed by the MIM method [39], but more study is warranted.

It is important to recognize that correlation is not equivalent to causal mechanisms. Causal relationship can only be established if it can be demonstrated that perturbation of the potential effector indeed has the predicted effect. To test whether a predicted DNA element is required for the establishment of epigenetic pattern, the most direct way is to genetically delete the element in question from the genome and then use a suitable experimental assay to evaluate whether the local epigenetic pattern changes as predicted. Recently developed genome-editing tools, such as TALEN [86] and CRISPR [87], are extremely useful for such validations.

CONCLUSIONS

Alignment-free methods have provided important computational tools for functional annotation of biological genome and for linking genome with epigenome. Conversely, the applications themselves have also provided important insights that can be used to develop more accurate and efficient methods. Close interactions between computational and experimental biologists will likely result in significant advances in both frontiers.

Key Points

- There has been a lot of progress in applications of alignment-free methods to study the link between the genome and epigenome.
- Feature space representation enables the detection of weakly associated sequence features.
- Alignment-free methods are useful for prediction of epigenomic patterns.
- Different alignment-free methods have different strengths and weaknesses.

FUNDING

This research was supported by the NIH grant HG005085 to GY.

References

1. The Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**(7414):57–74.
2. Altschul S, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3):403–10.
3. Lipman D, Pearson W. Rapid and sensitive protein similarity searches. *Science* 1985;**227**(4693):1435–41.
4. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009;**6**(11s): 22–32.
5. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet* 2010; **11**(7):476–86.
6. Miele A, Dekker J. Long-range chromosomal interactions and gene regulation. *Mol Biosyst* 2008;**4**(11):1046–57.
7. Vinga S, Almeida J. Alignment-free sequence comparison–a review. *Bioinformatics* 2003;**19**(4):513–23.
8. Lim SJ, Tan TW, Tong JC. Computational Epigenetics: the new scientific paradigm. *BioInformation* 2010;**4**(7):331–7.
9. Yuan GC. Linking genome to epigenome. *Wiley Interdiscip Rev Syst Biol Med* 2012;**4**(3):297–309.
10. Bernstein B, Meissner A, Lander E. The mammalian epigenome. Cell. 2007;**128**(4):669–81.
11. Kornberg R.D., Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 1999;**98**:285–94.
12. Jenuwein T, Allis C. Translating the histone code. *Science* 2001;**293**(5532):1074–80.
13. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;**16**(1):6–21.
14. Ren B, Robert F, Wyrick J, *et al*. Genome-wide location and function of DNA binding proteins. *Science* 2000; **290**(5500):2306–9.
15. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;**129**(4):823–37.
16. Mikkelsen T, Ku M, Jaffe D, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;**448**(7153):553–60.
17. Meissner A, Mikkelsen T, Gu H, *et al*. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**(7205):766–70.
18. Hansen K, Timp W, Bravo H, *et al*. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011;**43**(8):768–75.
19. Lister R, Pelizzola M, Dowen R, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**(7271):315–22.
20. Yuan GC, Liu Y, Dion M., *et al*. Genome-scale identification of nucleosome positions in *S.* cerevisiae. *Science* 2005; **309**(5734):626–30.
21. Di Gesù V, Lo Bosco G, Pinello L, *et al*. A multi-layer method to study genome-scale positions of nucleosomes. *Genomics* 2009;**93**(2):140–5.
22. Bock C, Lengauer T. Computational epigenetics. *Bioinformatics (Oxford, England)* 2008;**24**(1):1–10.
23. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 2010;**28**(8):817–25.
24. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**(3): 215–16.
25. Larson JL, Yuan G. Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. *BMC Bioinformatics* 2010;**11**:557.
26. Yu P, Xiao S, Xin X, *et al*. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. *Genome Res* 2013;**23**(2):352–64.
27. Kharchenko PV, Alekseyenko AA, Schwartz YB, *et al*. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* 2011;**471**(5):480–5.
28. Hoffman MM, Buske OJ, Wang J, *et al*. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012;**473**(6): 473–6.
29. Egger G, Liang G, Aparicio A, *et al*. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 2004; **429**(6990):457–63.
30. Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**(4):693–705.
31. Moazed D. Mechanisms for the inheritance of chromatin states. *Cell* 2011;**146**(4):510–8.
32. Struhl K, Segal E. Determinants of nucleosome positioning. *Nat Struct Mol Biol* 2013;**20**(3):267–73.
33. Lienert F, Wirbelauer C, Som I, *et al*. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* 2011;**43**(11):1091–7.
34. Yuan GC. Targeted recruitment of histone modifications in humans predicted by genomic sequences. *J Comput Biol* 2009;**16**(2)::341–55.
35. Pearson WR. An introduction to sequence similarity ('Homology') searching. In: *Current Protocols in Bioinformatics*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2013.
36. Peckham H, Thurman R, Fu Y, *et al*. Nucleosome positioning signals in genomic DNA. *Genome Res* 2007;**17**(8): 1170–7.
37. Yuan GC, Liu JS. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol* 2008;**4**(1): e13.
38. Kaplan N, Moore IK, Fondufe-Mittendorf Y, *et al*. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 2009;**458**(7236):362–7.
39. Pinello L, Lo Bosco G, Hanlon B, *et al*. A motif-independent metric for DNA sequence specificity. BMC Bioinformatics. 2011;**12**:408.
40. Pham TH, Ho TB, Tran DH, *et al*. *Prediction of Histone Modifications in DNA. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering* 2007;959–66.
41. Yang Y, Nephew K, Kim S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. *BMC Bioinformatics* 2009;**12(Suppl. 3)**:S15.
42. Zheng H, Wu H, Li J, *et al*. CpGIMethPred: computational model for predicting methylation status of CpG

islands in human genome. BMC Med Genomics. 2013; **6(Suppl. 1)**:S13.

43. Das R, Dimitrova N, Xuan Z, *et al*. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci USA* 2006;**103**(28):10713–16.
44. Dan G. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge, UK: Press Syndacate of the Univestity of Cambridge, 1997.
45. Burrows M, Wheeler DJ. *A Block-sorting Lossless Data Compression Algorithm*. Technical Report 124. Palo Alto, California, USA: Systems Research Center, 1994.
46. Abouelhoda M, Kurtz S, Ohlebusch E. The enhanced suffix array and its applications to genome analysis. *Algorithms in Bioinformatics*. Berlin Heidelberg: Springer, 2002;449–63.
47. Ferragina P, Manzini G. Opportunistic data structures with applications. In: *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* 2000. Redondo Beach, CA.
48. Heng L, Richard D. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**(14):1754–60.
49. Ben L, Cole T, Mihai P, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences. *Genome Biol* 2009;**10**(3):R25.
50. Apostolico A, Denas O. Fast algorithms for computing sequence distances by exhaustive substring composition. *Algorithms Mol Biol* 2008;**3**(13):1–9.
51. Soares I, Goios A, Amorim A. Sequence comparison alignment-free approach based on suffix tree and L-words frequency. *ScientificWorldJournal* 2012;**2012**:450124.
52. Apostolico A, Comin M, Parida L. Mining, compressing and classifying with extensible motifs. *Algorithms Mol Biol* 2006;**1**(1):4.
53. Chen Y, Blackwell TW, Chen J, *et al*. Integration of genome and chromatin structure with gene expression profiles to predict c-MYC recognition site binding and function. *PLoS Comput Biol* 2007;**3**(4):e63.
54. Rasheed F, Alshalalfa M, Alhajj R. Adapting machine learning technique for periodicity detection in nucleosomal locations in sequences. In: *Lecture Notes in Computer Science, IDEAL 2007*, Vol. 19. Berlin Heidelberg: Springer, 2007, 870–9.
55. Segal E, Fondufe-Mittendorf Y, Chen L, *et al*. A genomic code for nucleosome positioning. *Nature* 2006;**442**(7104): 772–8.
56. Field Y, Kaplan N, Fondufe-Mittendorf Y, *et al*. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* 2008;**4**(11): e1000216.
57. Barford LA, Fazzio RS, Smith DR. An Introduction to Wavelets. *Technical Report*. Bristol, UK: Hewlett-Packard Laboratories, 1992.
58. Cui K, Zang C, Roh T, *et al*. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* 2009;**4**(1):80–93.
59. The Encode Project Consortium Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;**447**(7416):799–816.
60. Levitsky V, Ponomarenko M, Ponomarenko J, *et al*. Nucleosomal DNA property database. *Bioinformatics* 1999; **15**(7–8):582–92.
61. Loshikhes L, Trifonov EN. Nucleosomal DNA sequence database. *Nucleic Acids Res* 1993;**21**(21):4857–9.
62. Shawe-Taylor J, Cristianini N. *Kernel Mehods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2006.
63. Kuksa P, Huang P, Pavlovic V. *Kernel Methods and Algorithms for General Sequence Analysis* 2008. http://paul.rutgers.edu.
64. Lodhi H, Saunders C, Shawe-Taylor J, *et al*. Text classification using string kernels. *J Mach Learn Res* 2002;**2**: 419–44.
65. Ming L, Xin C, Xin L, *et al*. The similarity metric. *IEEE Trans Inf Theory* 2004;**50**(12):3250–64.
66. Ferragina P, Giancarlo R, Greco V, *et al*. Compression-based classification of biological sequences and structures. *BMC Bioinformatics.* 2007;**8**:252.
67. Corona D, Di Benedetto V, Giancarlo R, *et al*. *The Chromatin Organization of an Eukaryotic Genome: Sequence Specific+Statistical=Combinatorial*. CoRR. 2012;abs/1205.6010.
68. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd edn. Boca Raton, Florida, USA: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1998.
69. Shawe-Taylor J, Cristianini N. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
70. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* 2011;**21**(12):2167–80.
71. Visel A, Blow M, Li Z, *et al*. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**(7231): 854–58.
72. Breiman L, Friedman J, Stone CJ, *et al*. *Classification and Regression Trees*. CRC Press, 1984.
73. Xie N, Liu Y. Review of decision trees. In: *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)* 2010. Chengdu: ;105–9.
74. Breiman L. Random forests. *Mach Learn* 2011;**45**(1):5–32.
75. Hugh AC, Edward IG, Robert EM. BART: Bayesian additive regression trees. *Ann Appl Stat* 2010;**4**(1):266–98.
76. Liu Y, Shao Z, Yuan GC. Prediction of Polycomb target genes in mouse embryonic stem cells. *Genomics* 2010;**96**(1): 17–26.
77. Lin WZ, Fang JA, Xiao X, *et al*. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PloS One* 2011;**6**(9):e24756.
78. Zhou Q, Liu JS. Extracting sequence features to predict protein–DNA interactions: a comparative study. *Nucleic Acids Res* 2008;**36**(12):4137–48.
79. Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo, California, USA: Morgan Kauffman, 1993.
80. Martin J, Hirschberg DS. *The Time Complexity of Decision Tree Induction. Technical Report 95 27*. California, USA: University of California at Irvine, 1995.
81. Belloni A, Chernozhukov V. On the computational complexity of MCMC-based estimators in large samples. *Ann Statist* 2009;**37**(4):2011–55.

82. Rouhani-Kalleh Algorithms for fast large scale data mining using logistic regression. In: *IEEE Symposium on Computational Intelligence and Data Mining, 2007*. Honolulu: CIDM, 2007.

83. Bordes A, Ertekin S, Weston J, *et al*. Fast Kernel classifiers with online and active learning. *J Mach Learning Res* 2005;**6**: 1579–619.

84. Chodavarapu RK, Feng S, Bernatavichute YV, *et al*. Relationship between nucleosome positioning and DNA methylation. *Nature* 2010;**466**(7304):388–92.

85. Hagarman JA, Motley MP, Kristjansdottir K. Coordinate Regulation of DNA Methylation and H3K27me3 in Mouse Embryonic Stem Cells. *PLoS One* 2013;**8**(1): e53880.

86. Miller J, Tan S, Qiao G, *et al*. A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* 2011;**29**(2): 143–8.

87. Mali P, Yang L, Esvelt K, *et al*. RNA-guided human genome engineering via Cas9. *Science* 2013;**339**(6121): 823–6.