BioSystems 138 (2015) 6–17

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Review Article

# The identification of *cis*-regulatory elements: A review from a machine learning perspective

Yifeng Li[a,b], Chih-yu Chen[a], Alice M. Kaye[a], Wyeth W. Wasserman[a,*]

[a] *Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia Vancouver, British Columbia V5Z 4H4, Canada*
[b] *Information and Communications Technologies, National Research Council of Canada, Ottawa, Ontario K1A 0R6, Canada*

ARTICLE INFO

*Article history:*
Received 18 June 2015
Received in revised form 9 October 2015
Accepted 14 October 2015
Available online 21 October 2015

*Keywords:*
*Cis*-regulatory elements
Gene regulation
Enhancers
Promoters
Machine learning
Deep learning
Ensemble learning
Data integration

ABSTRACT

The majority of the human genome consists of non-coding regions that have been called junk DNA. However, recent studies have unveiled that these regions contain *cis*-regulatory elements, such as promoters, enhancers, silencers, insulators, etc. These regulatory elements can play crucial roles in controlling gene expressions in specific cell types, conditions, and developmental stages. Disruption to these regions could contribute to phenotype changes. Precisely identifying regulatory elements is key to deciphering the mechanisms underlying transcriptional regulation. *Cis*-regulatory events are complex processes that involve chromatin accessibility, transcription factor binding, DNA methylation, histone modifications, and the interactions between them. The development of next-generation sequencing techniques has allowed us to capture these genomic features in depth. Applied analysis of genome sequences for clinical genetics has increased the urgency for detecting these regions. However, the complexity of *cis*-regulatory events and the deluge of sequencing data require accurate and efficient computational approaches, in particular, machine learning techniques. In this review, we describe machine learning approaches for predicting transcription factor binding sites, enhancers, and promoters, primarily driven by next-generation sequencing data. Data sources are provided in order to facilitate testing of novel methods. The purpose of this review is to attract computational experts and data scientists to advance this field.


**Contents**



* Corresponding author.
*E-mail addresses:* yifeng.li@nrc-cnrc.gc.ca (Y. Li), juliec@cmmt.ubc.ca (C.-y. Chen), akaye@cmmt.ubc.ca (A.M. Kaye), wyeth@cmmt.ubc.ca (W.W. Wasserman).

http://dx.doi.org/10.1016/j.biosystems.2015.10.002
0303-2647/

## 1. Introduction

In the human genome, less than 2% of the DNA sequence comprises protein-coding exons. The rest of the genome is non-coding and was previously regarded as junk DNA (Alexander et al., 2010). However, recent genome studies have unveiled that many of the non-coding sequences are transcribed and/or comprise regulatory regions used for transcriptional regulation (The ENCODE Project Consortium, 2012; Morris and Mattick, 2014). *Cis-regulatory elements* (CREs) are *cis*-acting non-coding DNA regions that regulate the transcription of genes. Promoters, enhancers, silencers, and insulators are among the key *cis*-regulatory elements (Fig. 1) (Noonan and McCallion, 2010). Within the nucleus of cells, active regulatory regions are nucleosome-depleted allowing *transcription factors* (TFs) to be recruited. Containing the *transcription start sites* (TSSs) of a gene, a promoter functions like a switch to turn on or off the transcription of the target gene (Fig. 1) (Lenhard et al., 2015). An enhancer (or silencer) can dynamically control the expression level of its target gene(s) through its interaction with promoters, even if they are far away from their target genes in the linear sequence space. An enhancer may reside in the intergenic region upstream or downstream of its target gene(s), and may also be embedded in an intronic region of a gene. Although distal to its target promoter(s) in linear space, a transcriptionally active enhancer is brought close to its target promoter by DNA looping in 3D nuclear space (Ong and Corces, 2011; Bickmore, 2013; Shlyueva et al., 2014) (Fig. 1). Two insulators can establish the boundaries of a regulatory domain within which an enhancer is unable to act beyond the insulator, blocking influence on the genes outside the domain (Fig. 1) (Raab and Kamakaka, 2010; Symmons et al., 2014; Liu et al., 2015). CREs play essential roles in determining which genes are specifically active in a cell type (Ong and Corces, 2012; Lovén et al., 2013; Hnisz et al., 2013), quantitatively controlling the expression levels of these genes at the right times, and confining the regulatory domains of certain functions (Symmons et al., 2014; Dowen et al., 2014). Variations in the *cis*-regulatory regions have been reported to cause assorted abnormal phenotype changes (Mathelier et al., 2015; Lupianez et al., 2015). Thus, identifying and annotating the CREs in the human genome is an important goal for clinical genetics.

Previously it was difficult to accurately annotate the non-coding regions due to the complexity of regulatory mechanisms and the lack of in depth data. Predictions of *transcription factor binding sites* (TFBSs) based purely on *position weight matrices* (PWMs) (Wasserman and Sandelin, 2004) have been useful to narrow down potential binding sites, but can suffer from high rates of false positives. In virtue of *next-generation sequencing* (NGS) techniques snapshotting various aspects of the genome, it becomes possible to identify CREs genome-wide. ChIP-seq (chromatin immunoprecipitation followed by sequencing) enables us to identify TFBSs and histone modifications (Johnson et al., 2007). RNA-seq techniques can precisely indicate the transcriptional activity of genes and exons (Wang et al., 2009). Nucleosome-depleted regions likely to contain CREs can be identified by DNase-seq (DNase I hypersensitive sites sequencing) (Boyle et al., 2008) and FAIRE-seq (formaldehyde-assisted isolation of regulatory elements) (Giresi et al., 2007). The chromatin interactions in 3D space can be captured by ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) (Fullwood et al., 2009) and Hi-C (high-throughput chromosome conformation capture) (Dixon et al., 2012) techniques. CAGE (cap analysis gene expression) (Andersson et al., 2014) and GRO-seq (global run-on and sequencing) (Core et al., 2008) are able to capture the TSSs of promoters and *enhancers RNAs* (eRNAs).

How can we best take advantage of the large volumes of genome-scale data generated by these techniques in order to pinpoint CREs across the entire genome? Machine learning consists of statistical modelling techniques that automatically learn useful knowledge from input data and infer unknowns based on a set of knowns. Thus, these data-driven intelligent algorithms emerge as key tools for the precise identification of CREs.

In this review, we focus on existing and potential machine learning approaches for the prediction of CREs by incorporating various genome-scale data sets. Instead of simply listing all related machine learning methods, the availability of training regions and the integration of various genomic data sets are the main focus throughout this review. Several reviews with different perspectives have been recently published. For a deeper discussion of the properties of enhancers, please refer to Pennacchio et al. (2013). Informative features to predict enhancers are reviewed in Shlyueva et al. (2014) and Wang et al. (2013) (this review also surveyed supervised methods). Our group has reviewed the methods of identifying TFBSs and predicting the impact of variations within TFBSs in Mathelier et al. (2015). See Lam et al. (2014) and Lai and Shiekhattar (2014) for the potential functionality and mechanisms of enhancer RNAs in gene transcription. Methods used in the pre-NGS era are reviewed in Wasserman and Sandelin (2004) and Pan (2006).

The rest of this review is organized as follows. The main sources of NGS data used for machine-learning based CRE predictions are given in Section 2. Unsupervised learning methods are reviewed in Section 3. We summarize supervised methods in Section 4. Section 5 covers deep learning methods. Future directions are discussed in Section 6.

## 2. Open-source data

Over the last few years, a tremendous amount of NGS data has been generated by several big consortia, each focusing on different goals (see Table 1). The ENCODE (Encyclopedia of DNA Elements) Consortium (The ENCODE Project Consortium, 2012) aims to build a comprehensive list of functional elements in the human genome. The goal of the NIH Roadmap Epigenomics Program (Roadmap Epigenomics Consortium, 2015) is to create an epigenomic atlas for primary cells and tissues in human. The objective of the FANTOM5 (Functional Annotation Of the Mammalian Genome) Project (The FANTOM Consortium, 2014; Andersson et al., 2014) is to uncover transcriptional regulatory networks based on transcript initiation positions. NGS data used in published articles are frequently deposited in the GEO (Gene Expression Omnibus) data