

A survey on computational methods for enhancer and enhancer target predictions

Qin Cao and Kevin Y. Yip

Department of Computer Science and Engineering, The Chinese University of Hong Kong,
Shatin, New Territories, Hong Kong
Tel: (852) 39438418; Fax: (852) 26035024
kevinyip@cse.cuhk.edu.hk

Enhancers are important cis-regulatory elements that play critical roles in a wide range of cellular processes by enhancing expression of target genes through promoter-enhancer loops. There are many interesting biological questions about enhancers, including their evolution and the relationships between their dysregulation and genetic diseases. The recent developments of experimental methods such as high-throughput reporter assays and ChIA-PET have enabled large-scale identification of enhancers and their targets. However, the current lists of identified enhancers and enhancer targets remain incomplete and unreliable due to the high noise level or low resolution of these methods. As a result, computational methods have emerged as an alternative for predicting the genomic locations of enhancers and their target genes. These methods have used a variety of features for predicting enhancers, including sequence motifs and epigenomic modifications. Potential enhancer targets have been predicted using activity correlations, distance constraints, and other features. Both prediction tasks are non-trivial due to cell-type specificity of enhancer activities, lack of definite orientation and distance of an enhancer from its target genes, insufficient known examples for training computational models, and other complexities. In this survey, we discuss the current computational methods for these two prediction tasks and analyze their pros and cons. We also point out obstacles of computational prediction of enhancers and enhancer targets in general, and suggest future research directions.

Introduction

Enhancers are important transcriptional regulatory DNA elements that can enhance transcription of target genes by recruiting transcription factors (TFs), which bring an enhancer close to the promoter of its target gene and trigger interactions with RNA polymerase II.

Strong sequence conservation at a non-coding region is a strong indicator of a potential enhancer (Pennacchio et al. 2006), especially when conservation is measured in ways related to the function, such as clustering or protein binding sites (Berman et al. 2004). Active enhancers are usually enriched in the histone mark H3K27ac, while both active and poised enhancers are enriched in H3K4me1, and latent enhancers lack these marks in general (Shlyueva et al. 2014). A typical enhancer is several hundred base pairs long as defined by transcription factor binding signals, while much longer enhancers called super-enhancers, which are bound by the Mediator complex and master transcription factors, have been found to be important in the control of cell identity (Whyte et al. 2013; Hnisz et al. 2013).

Previous studies have uncovered that enhancer dysregulation could cause abnormal gene expression and lead to genetic diseases (Carroll 2008; Visel et al. 2009; Dawson & Kouzarides 2012; Shlyueva et al. 2014), making enhancers an important study topic of both conceptual and practical values. Understanding the sophisticated operational mechanisms of enhancers has become a crucial part towards a complete understanding of the landscape of gene regulation.

In this chapter, we describe computational methods for identifying enhancers and their targets. We start with a brief introduction to current experimental approaches to these two tasks, based on which we discuss their main limitations and introduce computational methods as a key alternative. We then discuss the states-of-the-art in computational enhancer prediction, from the features used to both unsupervised and supervised methods. We next discuss computational methods for predicting enhancer-target promoter associations. Finally, we conclude the chapter and discuss future research directions on these two problems.

Introduction to current experimental approaches to testing enhancer activities and enhancer-target associations

Enhancers can be tested experimentally by different kinds of reporter assays (Shlyueva et al. 2014; ENCODE Project Consortium et al. 2012; Kwasnieski et al. 2014), including *in vivo* systems such as embryos of transgenic mice (Visel et al. 2007). To scale up reporter assays for testing many enhancers at the same time, high-throughput multiplexed reporter assays have been developed (Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012; Sharon et al. 2012). These methods have been applied to test previously predicted enhancers. For example, a recent study (Kwasnieski et al. 2014) has tested human enhancers predicted by the ENCODE consortium (ENCODE Project Consortium et al. 2012), and found that around 26% of these enhancer predictions have regulatory activities in the K562 cell line.

Another high-throughput method that can test the enhancer activities of millions of candidates simultaneously is STARR-seq (Arnold et al. 2013). The main novelty of this method is placing each enhancer to be tested downstream of the reporter gene, such that the enhancer sequence itself becomes part of the resulting RNA transcript. Standard RNA-sequencing (RNA-seq) can then be applied to measure quantitatively the activity of each enhancer by counting the number of reads containing the enhancer sequence.

A common limitation of these methods is that they do not preserve the whole native context of the predicted enhancers. For example, if an enhancer is predicted to be active in a context (cell/tissue type, development stage, disease state, etc.) but is tested in another context or even in another species, the chromatin state around the enhancer could be different, the TFs that bind the enhancer may not be expressed, and the genome structure required for enhancer-promoter looping could be altered. This means an enhancer that could be active in certain contexts may not show activities in a reporter assay, and even if it shows activities in a reporter assay, in which natural contexts it would be active is still unknown.

It is also important to note that these high-throughput experimental methods have been mainly used for testing enhancer candidates already defined by some other means, but not for discovering enhancers *ab initio*. In theory it should be possible to tile a major portion of a genome for testing the enhancer activities of the involved genomic regions using these high-throughput experimental methods. Such large-scale datasets are remained to be seen.

Many experimental approaches to enhancer-promoter association predictions rely on techniques that can capture chromosome conformations based on chromosome conformation capture (3C) (Dekker et al. 2002). There are many extended versions of 3C, such as circularized chromosome conformation capture (4C) (Zhao et al. 2006), chromosome conformation capture carbon copy (5C) (Dostie et al. 2006), genome-wide chromosome conformation capture (Hi-C) (Lieberman-Aiden et al. 2009) and chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) (Fullwood et al. 2009). Hi-C and ChIA-PET have facilitated whole-genome identification of DNA regions that are in close proximity in the three-dimensional genome structure but are not necessarily adjacent to each other in the primary DNA sequence, without requiring an input set of candidates. Among these two techniques, ChIA-PET further requires that a chosen factor, such as RNA polymerase II, is involved in the DNA contacts. If a promoter and a predicted enhancer are found to interact based on these chromosome conformation data, the promoter would be predicted as a target of the enhancer.

In order to study enhancer-promoter contacts, the chromosome conformation data need to have a very high (<10kb) resolution. Correspondingly, a large amount of sequencing data needs to be produced to ensure statistical stability at such a high data resolution, since the contact map matrix could be very sparse and unstable without sufficient data. Several recent studies have used Hi-C and ChIA-PET to study DNA contacts in human cell lines at sub-10kb resolutions (Jin et al. 2013; Heidari et al. 2014; Rao et al. 2014). These studies represent the current states-of-the-art in studying DNA long-range interactions.

While high-throughput chromosome conformation data have provided various insights about enhancer-promoter associations, they are still unable to comprehensively and accurately determine the targets of all enhancers for a number of reasons. First, having a physical interaction does not necessarily imply a functional relationship. In particular, many DNA contacts observed in Hi-C data may not be relevant to promoter-enhancer interactions (Shlyueva et al. 2014). Second, these high-throughput data could be noisy and are subject to different types of bias (DeMare et al. 2013; Duan et al. 2010; Li et al. 2010). Third, enhancer-promoter associations are also context-specific, and thus experimental data from a given context may not be relevant to other contexts.

Due to these limitations of current experimental approaches, the numbers of experimentally proven enhancers and enhancer-target associations are still limited, both in general and in particular contexts. As a result, computational methods have been widely used as an alternative in identifying enhancers and their targets. The advantage of using computational

methods is that they can utilize different types of available data to make predictions in an inexpensive way as compared to their experimental counterparts. In the past 15 years, many computational methods have been proposed, using ideas and data ever more advanced. The last few years have seen a rapid adaptation of high-throughput data originally generated not specifically for studying enhancers in these methods. As of today, both computational enhancer prediction and enhancer target prediction are still very active areas of research with new discoveries being constantly published.

Difficulties in computational predictions of enhancer and enhancer-promoter associations

Before going into the details of these computational methods, we first discuss the difficulties of the corresponding problems that explain the continuous need for better methods. These difficulties lie in several aspects, mainly related to the intrinsic properties of enhancers and the lack of high-confidence examples of experimentally validated enhancers and enhancer targets.

First, there is no simple rule governing the relative location of an enhancer from a gene that it targets. It can be positioned either upstream or downstream of the transcription start site (TSS) of its target gene. It can reside in an intergenic region, an intron, or even an exon of another gene. It can be as close as ten kilobases or as far as hundreds of kilobases or more from the target promoter. A recent study has suggested that the median distance between enhancers and their target promoters is 124kb (Jin et al. 2013). All these flexibility in enhancer location makes them much harder to identify than some other types of sequence elements, such as promoters, which are right upstream of the target genes.

Second, up to now, no single features or combinations of features have been found that can perfectly locate enhancers or determine enhancer-promoter associations (Shlyueva et al. 2014). The different features used by existing computational methods all have their pros and cons, which we will discuss in detail in the next section.

Third, enhancer activities and enhancer-promoter associations are both context specific. A recent study that analyzed data from twelve human cell lines has suggested that among the two, enhancer-promoter associations have relatively stronger cell type specificity (He et al. 2014). Context-specificity implies that computational methods using static features that do not change with the context, such as DNA sequence patterns, can only predict whether a genomic region could be an enhancer but not the contexts in which it is active, and only whether an enhancer could target a gene, but not the contexts in which the enhancer actually regulates the gene. This property implies that computational methods need to incorporate information from the context of interest in their predictions (Yip et al. 2013).

Fourth, enhancers and promoters could associate with each other in a multiple-to-multiple manner. In other words, one enhancer can target multiple promoters and one promoter can be targeted by multiple enhancers (He et al. 2014). As a result, some standard computational

methods that deal with one object at a time may not be suitable for predicting enhancers and enhancer targets.

Lastly, the lack of comprehensive lists of experimentally tested enhancers and enhancer targets means that there are limited examples for computational methods to reference. Some computational methods, especially those based on machine learning, require adequate positive and negative examples for modeling the general features of enhancers and enhancer targets. As a result, different studies have used a variety of ways to define “gold-standard” enhancers and enhancer targets for training their methods. A lot of these “gold-standard” examples are either not experimentally tested, or are taken from another context that may not be relevant to the context of interest. The devoid of experimentally tested examples also means that computational predictions cannot be easily validated without performing additional experiments.

Owing to all these difficulties, computational methods should be considered a supplement to experimental methods rather than a replacement. Computational predictions of existing methods all need to be experimentally tested to confirm their correctness.

Computational methods for enhancer prediction

The problem of computational prediction of enhancers is defined as follows. Given a set of genomic regions, each of which is described by a set of features, the goal is to identify the regions that correspond to enhancers based on the features.

This definition requires an input list of genomic regions the status of which (enhancer or non-enhancer) is to be predicted. In many cases, one only wants to predict an approximate location of each enhancer, in which case it is common to divide the whole genome into bins of a fixed size, and predict whether each bin overlaps an enhancer or not. On the other hand, if the predicted enhancers are to be tested experimentally, it is necessary to make sure that an enhancer candidate includes the core part of the enhancer, such as the TF binding sites (TFBSs). In this scenario, the raw predictions need to be further refined.

Many computational methods have been proposed for this prediction task. They differ from each other by the features they use and the way the features are used to make the predictions. In the followings, we first describe the features considered by different enhancer prediction methods, and then move on to discuss these methods themselves.

Features used in enhancer prediction

Many types of features have been considered in predicting enhancers (Table 1 and Figure 1). Before the boom of high-throughput sequencing data that probe different types of features related to enhancers in a context-specific manner, researchers predicted cis-regulatory

modules (CRMs), enhancers included, largely based on evolutionary conservation and sequence motifs (Su et al. 2010). Evolutionary conservation signifies regions with functional importance. Non-coding regions, including intergenic regions and introns, with unexpectedly strong evolutionary conservation could be CRMs. On the other hand, some functionally conserved enhancers do not have high sequence conservation (Su et al. 2010; Meireles-Filho & Stark 2009). This could indicate that conservation is not sufficient for identifying enhancers, or that the way to measure conservation needs to be improved (Berman et al. 2004).

Regions with a good match to a sequence motif could be binding sites of the TF. Excluding binding sites at annotated regions such as promoters, the remaining could be CRMs, especially for regions with a high density of motif matches (Su et al. 2010). Since TF binding also depends on factors other than the sequence, sequence motifs can be considered a weak feature for enhancer prediction.

As discussed above, using these static features to predict enhancers could at best identify regions with a potential to be an enhancer, without telling the contexts in which the enhancers are actually active. It is also hard to use conservation and sequence motifs alone to distinguish enhancers from other types of regulatory elements such as silencers and insulators without a thorough understanding of the factors that bind these different types of elements.

Later on, the development of ChIP-seq (Park 2009) made it easy to measure DNA-binding affinity of transcription factors genome-wide (Bailey & MacHanick 2012). Compared to sequence motifs, the TFBSs identified by ChIP-seq are directly measured in the context of interest. They were thus used to predict enhancers in a context specific manner (Yip et al. 2012). Again, some of these binding sites may correspond to other types of functional enhancers (Shlyueva et al. 2014; Li et al. 2008). Moreover, there is a limited number of TFs with ChIP-seq data available, making it impossible to rely on ChIP-seq data alone to identify all TFBSs for enhancer prediction. Standard ChIP-seq data also have limited data resolution. This problem has been tackled by a new method called ChIP-exo (Rhee & Pugh 2011; Rhee & Pugh 2012), which provides close to single nucleotide precision of TFBSs by enzymatically digesting unbound portions of the pulled-down DNA.

ChIP-seq experiments were also used extensively in studying various types of histone modifications (HMs) at whole-genome scales. Some HMs were found to be highly related to enhancers, including H3K4me1 that marks both poised and active enhancers, and H3K27ac that marks active enhancers (Rada-Iglesias et al. 2011). These HMs provide a way to distinguish enhancers from other types of regulatory elements, such as promoters, which are marked by H3K4me3. On the other hand, while H3K4me1 and H3K27ac have been well-recognized as important enhancer marks, there has not been a consensus as to whether they are sufficient or necessary for identifying active enhancers. For instance, a recent study has found that H3K4me3 (as a negative feature for enhancers), H3K4me1 and H3K4me2 are the top three HMs for enhancer prediction while H3K27ac was not selected as one of the most important predictors (Rajagopal et al. 2013), although H3K27ac is widely used in many other

studies as an indicator of active enhancers. Some previous studies have also shown that no single types of HM or a combination of several HMs could predict enhancers perfectly (Arnold et al. 2013), and some active enhancers do not have typical active marks (Bonn et al. 2012). Despite these complications, HMs still represent a cost-effective set of features in identifying context-specific enhancers, in that only a small set of ChIP-seq experiments are sufficient for identifying a fairly accurate set of active enhancers in a context.

Pushing this idea further is to use one single context-specific feature in identifying enhancers. One popular choice is chromatin accessibility as measured by DNase I hypersensitivity using DNase-Seq (Boyle et al. 2008) or FAIRE-Seq (Giresi et al. 2007). These data indicate genomic regions with high accessibility of chromatin where DNA sequences are depleted of nucleosomes, which signify functional activities of these regions. Active enhancers were found to overlap with DNase hypersensitive sites (DHSs), but obviously not all highly accessible genomic regions are enhancers. Chromatin accessibility data can thus be used to limit the search space of active enhancers to only the DHSs, and let the precise locations be identified with the help of other features such as TF sequence motifs.

Recently, it has been discovered that active enhancers produce short (<2kb) non-coding RNAs called eRNAs in a bi-directional manner (Kim et al. 2010). Based on this idea, a recent study has identified enhancers as regions with some bi-directional transcription patterns (Andersson et al. 2014), according to the abundant CAGE-based TSS data from FANTOM5 (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014). Currently, knowledge about eRNAs, including their functional mechanisms, is still limited. It is not yet clear whether active enhancers must produce eRNAs, and whether genomic regions producing eRNA-like RNAs must be enhancers. Nevertheless, the idea of using eRNAs to identify enhancers has become popular due to the wide availability of transcriptome data.

Less popularly, some studies have attempted to use DNA methylation level to predict enhancers (Aran et al. 2013). The role of DNA methylation in marking repressed promoters has long been recognized. Many inactive genomic regions are also marked by DNA methylation. Due to the diverse types of regions marked by DNA methylation, data about DNA methylation in a single context can hardly be used to identify enhancers. However, if two contexts are being compared (e.g., tumor vs. normal tissue), sites with differential DNA methylation could have differential activities in the two contexts, and some of them could correspond to functional elements such as enhancers. Currently, the degree of enhancer activities reflected by their DNA methylation levels is still unknown. The roles of different types of DNA methylation, such as 5-mC and 5-hmC (Xu et al. 2011), in regulating enhancer activities are also unclear.

Some studies have used correlation information between enhancer candidates and promoters to predict enhancers (Thurman et al. 2012). The main idea is that some activity indicators of enhancers (such as H3K27ac) are believed to correlate strongly with the transcription of their target genes across multiple contexts. If a non-promoter genomic region is found to exhibit such a correlation with a gene, the region could be an enhancer that regulates the gene. This

idea is also commonly used in identifying enhancer targets. It has some limitations as we will discuss later.

Table 1 summarizes the features used in current computational methods for enhancer predictions discussed above. A detailed discussion on the pros and cons of some of these features in identifying enhancers from the perspective of biological experiments can be found in a recently published review (Shlyueva et al. 2014).

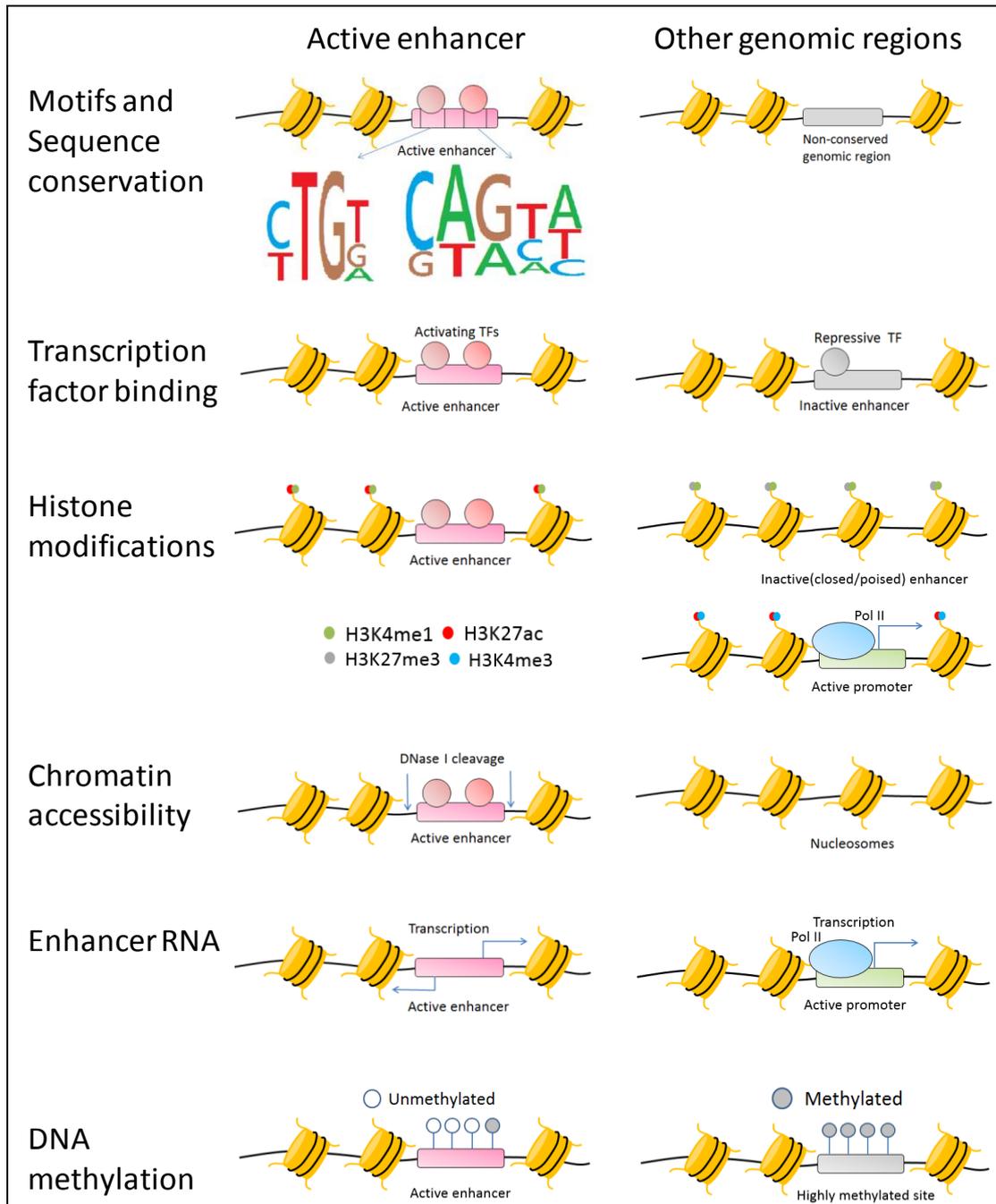
It should be noted that there are some additional enhancer features that have been more commonly used to define “gold-standard” enhancers instead of being used in the prediction process. For example, previous studies have shown that a large fraction of the binding sites of transcriptional co-activator proteins P300 and CBP are enhancers (Blow et al. 2010; May et al. 2011; Ramos et al. 2010). As a result, they have been used in some studies to define gold-standard enhancers (He et al. 2014; Rajagopal et al. 2013). One likely reason that binding sites of these proteins have not been as popularly used as enhancer predictors is that they are found in only a subset of active enhancers. This means although their presence stronger indicates an enhancer, using them as the only features could lead to a lot of false negatives.

Table 1. A summary of features used in computational enhancer prediction

Feature	Advantages	Potential drawbacks
TF binding motifs	Widely available	Presence of a motif does not guarantee binding of a TF in a given context; A TF could bind regions without a canonical sequence motif; Many TFBSs are not within enhancers
Evolutionary conservation	Widely available	Some functional enhancers do not have high sequence-level conservation; Cannot distinguish between different types of conserved DNA elements; Does not provide context-specific information
TFBSs based on ChIP-seq or ChIP-exo	Directly measured from the context of interest	Many TFBSs are not within enhancers; Requires many ChIP-seq experiments to obtain a comprehensive list of binding sites for many TFs
HMs	Provides information about both poised and active enhancers; There are both positive and negative HMs for enhancers;	No single HMs or their combinations have been found to correlate perfectly with enhancer activities

	Only a small number of ChIP-seq experiments is needed for each context	
Chromatin accessibility	Only a single type of features is required for each context	Regions with high chromatin accessibility do not necessarily correspond to enhancers
eRNA	One of the most accurate single features for enhancer prediction; Transcriptome data are widely available	The detailed mechanisms of eRNA remain to be explored; Active enhancers may not produce eRNAs; Regions producing eRNA-like RNAs may not be enhancers; Many produced RNA-seq data are poly-A enriched, which may not contain eRNA signals
DNA methylation	Provides complementary information to the other features	Quantitative relationship between enhancer methylation and target gene expression is still unclear; Different types of DNA methylation may play different roles in enhancer regulation

Figure 1. Features used in computational enhancer prediction. The left part of the figure shows features of active enhancers while the right part shows the corresponding features of inactive enhancers, other regulatory elements (such as promoters), or other genomic regions.



Unsupervised methods for enhancer prediction

Many computational methods have been proposed for predicting enhancers using the features described above. Traditional methods that mainly use non-context-specific features have been discussed in detail in another review (Su et al. 2010). Here we focus on more recent methods that incorporate the different types of context-specific features. These methods can be broadly grouped into two categories, namely unsupervised methods and supervised methods.

Unsupervised methods do not require any known enhancers and non-enhancers as examples. Some of these methods define simple filtering rules to identify the most likely enhancers

based on the observed features. Some other methods cluster genomic regions according to these features, and identify clusters that are likely enhancers. In contrast, supervised methods require known enhancers and non-enhancers as inputs and derive models for enhancers using machine learning techniques. A more detailed discussion of the use of unsupervised and supervised methods (and also semi-supervised methods) in identifying genomic elements can be found in a recent review (Yip et al. 2013). In this section we first discuss the unsupervised methods for enhancer prediction.

Thurman et al. (Thurman et al. 2012) defined distal DHSs separated from a TSS by at least one other DHS as enhancer candidates. The DNase I hypersensitivity signals of each enhancer candidate in different cell types were correlated with those of each promoter within 500 kb from it. Any candidate with a resulting Pearson correlation of 0.7 or above was predicted as an enhancer.

Andersson et al. (Andersson et al. 2014) identified enhancers based on a directionality score of eRNA. They superimposed CAGE tags on H3K27ac-marked enhancers defined by ENCODE (the methods of which will be discussed below) and found that CAGE tags showed a bimodal distribution flanking the central P300 peak with divergent transcription from the enhancer. In contrast, the transcripts at promoters were strongly biased towards the sense direction. With this distinct difference, a directionality score was calculated for every 200bp window genome-wide, and loci with low, non-promoter-like directionality scores were selected as enhancer candidates, among which the ones located far away from TSSs and exons of protein-coding and non-coding genes were predicted as enhancers. To validate these predicted enhancers, they further selected strong, moderate and low-activity enhancers defined by CAGE tag frequency in HeLa cells and conducted enhancer reporter assays. They found that 73.9%, 70.7% and 67.4% of the strong, moderate and low-activity CAGE-defined enhancers showed significant signals in the reporter assays, respectively, demonstrating that eRNA could be an intrinsic characteristic of active enhancers.

The above two methods are simple unsupervised methods based on thresholding on a single feature. ChromHMM (Ernst & Kellis 2010; Ernst & Kellis 2012; Ernst et al. 2011) and Segway (Hoffman et al. 2012) utilize more complex machine learning models and dozens of features each in predicting enhancers.

ChromHMM characterizes chromatin states including enhancers by learning a multivariate hidden Markov model (HMM) with the largest data set available at the time it was proposed (Barski et al. 2007; Wang et al. 2008) containing various HMs, histone variants and protein binding ChIP-seq signals (e.g. H2AZ, RNA polymerase II and CTCF) (Ernst & Kellis 2010). This method involves five key steps. First, the whole genome was divided into 200bp intervals. The signals of different HMs in an interval were then binarized, and thus each interval was described by a binary vector of the presence/absence of HM signals. Third, the number of states and the model parameters were determined by an exhaustive comparison of the cluster number from 2 to 80, with 3 different types of random initialization of parameters. The best model was selected by a Bayesian Information Criterion (BIC) score. Intuitively, the

procedure attempted to find the minimum number of states that could still distinguish genomic regions exhibiting distinct HM patterns into different states. Finally, a 51-state model was selected. The fourth step was to associate each genomic interval with the state that maximizes the posterior probability using the forward-backward algorithm. The last step was to interpret the states biologically. This step involved both analyses of additional data (including expression, sequence motif, gene ontology, SNP and GWAS) and manual annotations. Based on the annotation results, several states were found to be related to enhancers (States 20, 29, 30, 31, 32 and 33). For instance, genomic regions in States 29 and 30 were interpreted as strong distal enhancers with characteristic high DNase I hypersensitivity and TF binding signals.

The same authors later **Error! Reference source not found.** further applied ChromHMM to nine human cell types and identified 15 states that showed distinct enrichments of different types of biological signals (Ernst et al. 2011). Eight predicted strong enhancers (State 4) and 7 predicted weak/poised enhancers (State 7) from the Hep-G2 cell line and 7 predicted weak/poised enhancers specific to the GM12878 cell line were tested in Hep-G2 using luciferase reporter assays. Only strong enhancers from HepG2 were observed to show strong luciferase signals.

Segway, based on Dynamic Bayesian Network (DBN), is similar to ChromHMM in the underlying mechanisms. In fact, a standard HMM can be represented by a DBN (Koller & Friedman 2009). The main differences between the original applications of Segway and ChromHMM lie in the following aspects: First, Segway used HMs and TF binding as features while ChromHMM mainly used HMs; Second, Segway worked at single base pair resolution while ChromHMM worked on 200bp bins; Third, Segway accepted continuous features while ChromHMM dealt with binary features; Fourth, Segway had an explicit indicator variable for missing values while ChromHMM considers them as 0s. The first two differences were mainly choices made in the corresponding studies, but the ChromHMM method itself could incorporate TF binding signals and work at a higher resolution. When applying to the dataset from ENCODE, Segway identified 25 labels (analogous to the “states” in ChromHMM) and marked enhancers by the E-label. In a later study, the authors of ChromHMM and Segway collaborated and integrated these two methods to identify sequence elements from ENCODE data (Hoffman et al. 2013).

Yip et al. (Yip et al. 2012) defined two pipelines for predicting enhancers. Both pipelines start from all genome regions, and apply a series of filters to retain only regions likely to be enhancers. The first pipeline involves ChIP-seq signal shapes, gene annotations and HM signals. The second pipeline involves sequence features, TF binding active regions (BARs), gene annotations, conservation scores, sequence motifs and TF expression levels. BARs were determined using ENCODE TF binding data. Although ChIP-seq data of more than 100 TRFs were collected, this number of TRFs is still only a small portion of the estimated 1,700 to 1,900 human TFs (Vaquerizas et al. 2009). Therefore, instead of defining BARs by the binding sites directly observed in the limited data, a statistical model of BARs was

constructed using these directly observed binding regions as positive examples and various types of ENCODE data as the features, including DNase I hypersensitivity and HMs.

Predictions of the two pipelines were combined, and the integrated predictions underwent two rounds of experimental validations. In the first round, among 6 predictions randomly selected from the top 50 predictions, 5 were found to have enhancer activities in various tissues in mouse embryo with good reproducibility. In the second round, the goal was to predict all enhancers in the human genome. Therefore a large number of predictions were made, among which about 50 were experimentally tested in mouse and Medaka fish. Overall, 42 unique regions could be successfully tested, among which 28 showed enhancer activities in at least one assay.

Overall, the five methods described above represent some of the latest unsupervised methods for computational enhancer prediction. It should be noted that the first two methods were specially designed for enhancer prediction while the other three were designed to discover various types of chromatin states in general, but with enhancers as some of the states in particular.

Supervised methods for enhancer prediction

As explained, supervised methods for enhancer prediction require known enhancers and non-enhancers as input examples. Since the numbers of experimentally tested positive and negative examples are limited, different methods have used a variety of strategies to define these input examples. The different methods also differ from each other by the features being used and the statistical models constructed.

Heintzman et al. (Heintzman et al. 2007) used a correlation-based method to predict enhancers based on their similarity to the enhancer examples. Enhancer examples were defined as regions with P300 binding sites. The genome was divided into 10kb windows, where an HM profile was constructed for each window based on the average ChIP-seq signals of different HMs. Enhancers were then predicted as those windows having an HM profile highly correlated with a P300-defined enhancer. In total, around 700 enhancers were predicted in this way. They were found to be significantly enriched in predicted transcriptional regulatory modules and DHSs. A large fraction of these predictions were also found to contain highly conserved sequences.

Won et al. (Won et al. 2008) presented a HMM-based method integrating HMs to predict enhancers. The positive examples were again defined by P300 binding sites. A simulated annealing procedure was used to search for the most informative combination of HMs and the optimal window size. The procedure identified a set of 6 HMs as the most informative, and a window size of 2kb to be optimal. A 3-state HMM model was then trained on a subset of the enhancer examples, and tested on another subset. The prediction results were found to be

more accurate than the predictions by the Heintzman et al. method (Heintzman et al. 2007) in terms of positive predictive value and sensitivity.

Firpi et al. (Firpi et al. 2010) developed a method called CSI-ANN based on a time-delayed neural network (TDNN) framework to predict enhancers in HeLa and Human CD4⁺ T cells. In the case of T cells, the whole genome was divided into 2.5kb windows with consecutive windows overlapped by 1.25kb. Windows that contain gene-distal and narrow P300 binding peaks in human T cells and overlap computationally predicted enhancers in the PReMod database (Ferretti et al. 2007) were defined as enhancer examples, leading to a positive set of 213 enhancers. The negative set was composed of random windows 10 times the number of positive examples. For each window, the average signals of 39 HMs in T cells, or an energy function of them (D'Alessandro et al. 2003) were computed as its features. Fisher discriminant analysis (FDA) was then performed to reduce these $39 \times 2 = 78$ features to a one-dimensional feature. Finally this feature was fed into a TDNN classifier. 36,769 predictions were made and 13.1% of them were found to overlap P300 sites and DHSs in T cells. 22.1% of the predictions were found to be conserved across 17 vertebrate genomes and 24.6% were enriched for TF binding motifs.

Rajagopal et al. (Rajagopal et al. 2013) developed a vector-random-forest-based supervised model called RF ECS for enhancer prediction. Gene-distal P300 binding sites overlapping DHS were defined as positive enhancer examples, while TSSs overlapping DHS and random 100bp bins distal from P300 binding sites or TSS were defined as negative enhancer examples. For each 100bp genomic region, the average signal of each of 24 HMs was computed. However, instead of taking only these average signals as the features of a genomic region like what was commonly done, each region also took the signal values from the adjacent regions within the 1kb upstream and downstream window as its own features. Therefore for each genomic region, each HM produced a 20-dimensional feature vector of numeric values. The reason for doing that was to capture the local signal pattern, which could be useful for identifying enhancers. To handle these vector features, RF ECS constructed a linear classifier using the Fisher Discriminant approach inside each decision tree node.

This method was applied to the H1 embryonic stem cells and the IMR90 lung fibroblasts. To validate the predictions, some “gold standard” enhancer regions were defined by combining DHS, P300 binding sites and a few sequence specific transcription factors known to function in each of these two cell types. The validation rate of the predicted enhancers was 80% in H1, which was highly significant when compared to the 18.43% validation rate of randomly predicted enhancers. 5% of the predicted enhancers overlapped with TSSs, which were considered misclassified. The validation and misclassification rates in IMR90 were 85% and 4%, respectively. It should be noted that since the criteria used for defining the enhancer examples in the training set and the criteria used to define the validation set were not mutually exclusive, the accuracy of the model needs to be further confirmed by independent data sets.

Another contribution of this work was its proposed set of HMs optimal for enhancer predictions. The top three HMs were found to be H3K4me3, H3K4me1 and H3K4me2 in H1, while H3K27ac, commonly believed to mark active enhancers, seemed not very predicative.

In summary, due to the increasing number of experimentally validated enhancers and the availability of high-throughput features, supervised methods have become increasingly popular. It is expected that more supervised enhancer prediction methods will be proposed in the coming years.

Computational methods for enhancer target prediction

Features used in enhancer target prediction

Compared to enhancer prediction, less feature types have been considered in predicting enhancer targets (Table 2). The first and simplest feature considered is whether a promoter is the nearest one from an enhancer. A slight variation of this idea is to consider the distance between an enhancer and a promoter, assuming a higher possibility that the enhancer regulates a promoter if they are closer to each other. Some previous studies have considered enhancers between 125kb (Ernst et al. 2011) and 1Mb (Fu et al. 2014) from potential target promoters. As discussed, chromosome conformation data have suggested that the median distance between an enhancer and a target promoter is 124kb (Jin et al. 2013). One drawback of using distance to predict enhancer targets is that very distal associations could be missed if the distance threshold is set too low. Conversely, if the distance threshold is set too high, many false positives could be produced. One way to avoid setting an arbitrary distance threshold is to consider only enhancer-promoter pairs within same topologically associating domains (TAD) (Dixon et al. 2012; Nora et al. 2012), which are genomic blocks separated from other blocks by the genome structure.

Sequence co-conservation is another feature that has been used in enhancer-promoter association prediction (He et al. 2014). The rationale is that if an enhancer regulates a promoter, there would be selective pressure against independent evolution of them, and thus they may exhibit co-conservation patterns. Some previous studies (Ahituv et al. 2005; Kikuta et al. 2007) also suggested that a real enhancer-promoter association is more likely to be maintained in a conserved synteny block (Larkin et al. 2009), which could be used as a soft distance constraint.

As high-throughput sequencing data became widely available, the correlations between certain molecular signals at an enhancer and its candidate target promoters across multiple contexts were considered. As discussed above, the main idea is that if the activity of an enhancer correlates with the activity of a promoter, the enhancer could be regulating the promoter. The molecular signals considered and the potential issues of using correlation features have been discussed above when discussing the features used in enhancer prediction. An additional issue is that if correlations are computed between all enhancer-promoter pairs

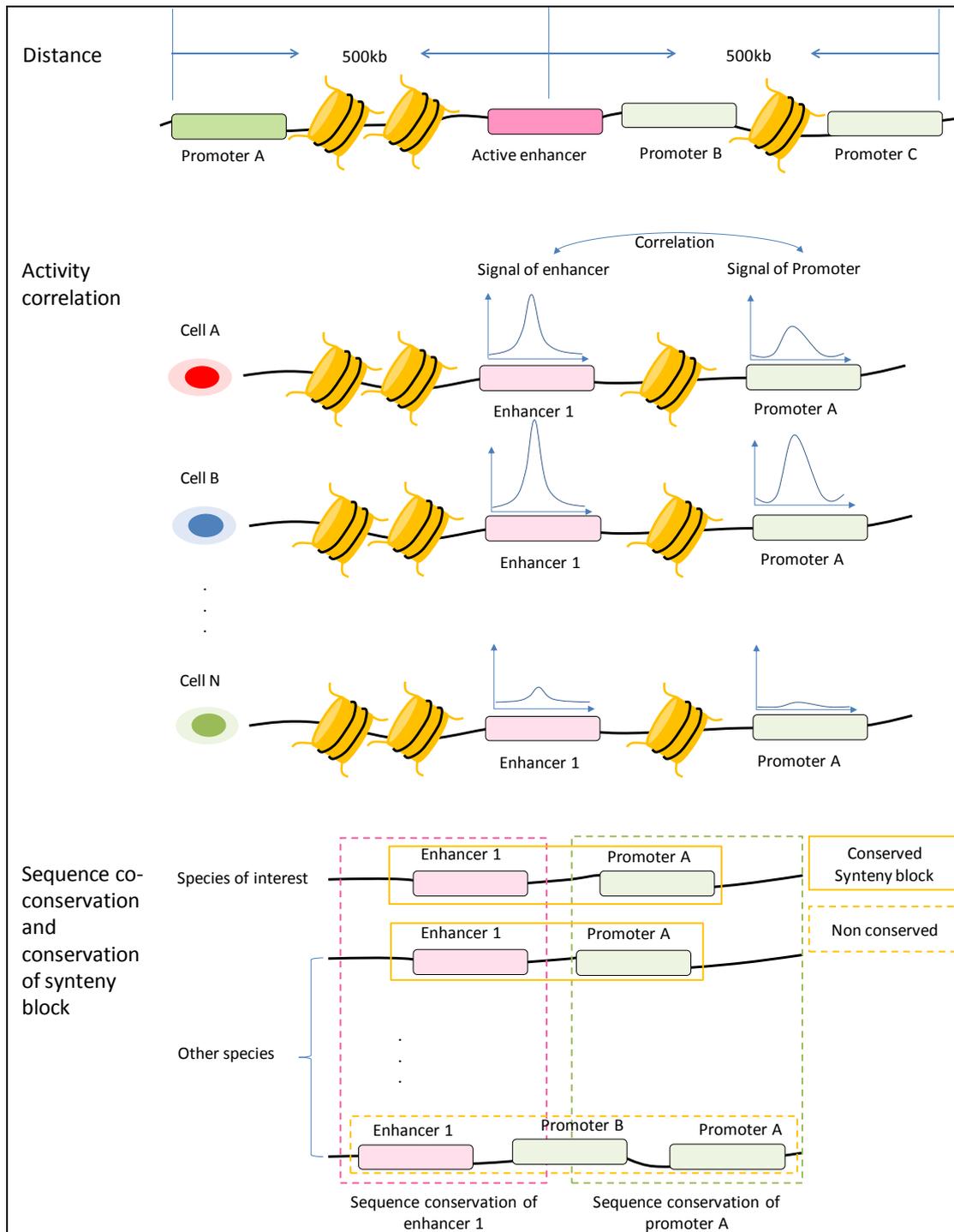
without any pre-filtering, there would be a very large number of pairs being considered. As a result, a very large number of contexts are needed to reach statistical significance after considering the issue of multiple hypothesis testing. We also note that to what extent enhancer activities can quantitatively correlate with promoter activities is still not clear. In fact, some studies (Andersson et al. 2014) have observed enhancer-promoter associations with low activity correlations.

Among these features, only signal correlations consider context-specific information. A tricky point is that depending on how this feature is used, it may still be unable to identify context-specific enhancer targets. For instance, if a single correlation value is computed based on all the contexts, this correlation value only tells whether the enhancer appears to regulate the promoter in general, but not exactly the contexts in which the regulation happens.

Table 2. A summary of features used in computational enhancer target prediction

Feature	Advantages	Potential drawbacks
Closest promoter	Easy to identify	An enhancer does not always regulate the closest promoter (Andersson et al. 2014; He et al. 2014)
Distance between enhancer and promoter	Easy to compute	There may not be a single threshold suitable for all cases; An enhancer does not always regulate the closest promoters
Co-conservation	Easy to compute; Utilizes information from multiple species	Both enhancers and enhancer-promoter associations are not necessarily highly conserved
Correlation of molecular signals	Utilizes context-specific information	No signal correlates perfectly between enhancers and promoters; Correlation coefficients could be strongly affected by outliers; Requires a large number of context to reach statistical significance

Figure 2. Features used in computational enhancer target prediction.



Unsupervised methods for enhancer target prediction

Similar to enhancer prediction, most methods for enhancer target predictions are unsupervised, due to the limited number of experimentally validated enhancers and enhancer targets.

As discussed, the most straightforward method is to predict the closest promoter as the only target of each enhancer. This is a simple but imperfect method. Several studies (Andersson et

al. 2014; He et al. 2014) have shown that only a fraction (e.g., 40% (Andersson et al. 2014)) of enhancers recognize the nearest promoter as their targets, and one enhancer could regulate multiple promoters. A variation of this method is to predict the nearest promoter within a certain distance range (e.g. between 5kb and 50kb from the enhancer (Ernst et al. 2011)) from an enhancer as its target.

Most current unsupervised methods extract all promoters within a certain distance range from an enhancer as candidate targets, and then use activity correlations to identify the most likely targets. A practical problem is finding a proper correlation threshold. Some studies (Andersson et al. 2014) use a rather low threshold of 0.2 while some other studies (Thurman et al. 2012) use a much higher value of 0.7. If a value-based correlation function such as Pearson correlation is used, the correlation values can be easily affected by a few outlier points. On the other hand, if a rank-based correlation function such as Spearman correlation is used, the correlation value can become quite arbitrary if the activity values in many contexts are similar and their ranks are sensitive to small differences. Multiple hypothesis testing is also a critical issue, because without a proper distance cutoff, many enhancer-promoter pairs would be considered and it is easy to get some strong correlation values merely by chance. Table 3 compares some of these unsupervised methods.

Table 3. A summary of correlation-based unsupervised methods for enhancer target prediction

Reference	Distance	Features denoting activity/inactivity (A/I: enhancer-promoter)	Correlation function(s)	Threshold
(Thurman et al. 2012)	Within 500kb	A: DNase I hypersensitivity -DNase I hypersensitivity	Pearson	0.7
(Andersson et al. 2014)	Within 500kb	A: CAGE(eRNA)-CAGE(mRNA)	Pearson	0.2
(Fu et al. 2014)	Within 1Mb	A: H3K4me1/H3K27ac-mRNA I: DNA methylation-mRNA	Pearson and Spearman	User-defined

Some studies used a further step to validate their predictions. Thurman et al. (Thurman et al. 2012) profiled chromatin interactions using 5C for the phenylalanine hydroxylase (PAH) gene in hepatic cell and found the chromatin interactions measured by 5C closely paralleled the correlations of the corresponding predicted associations. They also overlapped their predictions with 5C and ChIA-PET data in K562, and discovered that their predictions were markedly enriched in the DNA long-range interactions. Andersson et al. (Andersson et al. 2014) found that 15.3% of their predictions could be validated by ChIA-PET data from multiple cell types. Moreover, their predictions were enriched in conserved sequence motifs and ChIP-seq peaks.

Ernst et al. (Ernst et al. 2011) selected all TSSs between 5kb and 125kb from an enhancer as its potential targets. To identify the more likely ones, these enhancer-TSS pairs were first assumed to be the positive examples, and a set of negative examples was formed by randomly assigning expression values of the same pairs. For each (positive or negative) enhancer-TSS

pair, the correlation between the HM signals at the enhancer and the expression levels of the TSS across multiple contexts was computed. A logistic regression classifier was then constructed to distinguish the positive and negative examples based on the activity correlations. The classifier was then used to compute a link score for each enhancer-TSS pair, defined as the ratio of the positive association probability to the negative association probability. The pairs with a link score larger than 2.5 were predicted as real associations. This is another example that even a supervised model (logistic regression) was used, but since the positive examples were not really known examples but just a set of examples more likely to be positive due to the proximity of the corresponding enhancers and TSSs, the overall method should be considered an unsupervised one for predicting enhancer targets.

Corradin et al. (Corradin et al. 2014) developed a method called PreSTIGE for cell-type specific enhancer-promoter association prediction. Enhancers were defined as H3K4me1 sites across 12 cell types. First, a specificity score was assigned to each enhancer and to each transcript separately in the 12 cell types based on Shannon's entropy (Schug et al. 2005). Thresholds were set to define cell-type specific enhancers and transcripts based on the specificity scores. For example, enhancers with high specificity to a certain cell type were considered to be active in this cell type but not in the others. The next step was to link cell-type specific enhancers to their target cell-type specific genes. Several linear domain models for setting the distance thresholds were compared, based on which a model called 100kb/CTCF was selected to link enhancers and genes. In this model, all TSSs closer to an enhancer than the closest CTCF binding site, or 100kb at most, were predicted as the targets of the enhancer. This model identified over 226,000 and 113,000 enhancer-target predictions across the 12 cell types with low and high thresholds, respectively. The predictions were further overlapped with existing 3C, ChIA-PET, eQTL, 5C and colon cancer specific enhancer alteration data and showed significant intersections.

Supervised methods for enhancer target prediction

There have not been a lot of supervised methods proposed for enhancer target prediction, due to the limited number of validated examples. In this section, we introduce one supervised method that uses chromosome conformation data to define the examples.

A sophisticated Random Forest based supervised method called IM-PET was developed by He et al. (He et al. 2014). The positive examples were selected from enhancer-promoter pairs with ChIA-PET connections in K562 and MCF-7 cells, with the additional requirements that there were at least 5 PET counts, at least one of the two interacting sites contained P300 binding, and the other contained a promoter of RPKM larger than 0. A naïve way to define the negative examples would be to draw random enhancer-promoter pairs. However, if the promoters in these pairs were very far away from their enhancers, which would likely be the case if enhancers and promoters were drawn uniformly from the whole genome or the same chromosome, the positive and negative examples could be easily separated by a simple model that considers only the distance between the enhancer and promoter. Therefore, IM-PET

instead used random enhancer-promoter pairs with a distance that follows a background distribution of non-interacting genomic loci in a chromatin fiber (Dekker et al. 2002). The negative examples were also required not to have 3 or more PET counts in the ChIA-PET data. Four features were then used to train a supervised Random Forest model for enhancer-target associations. The first feature was the activity correlation between an enhancer and a promoter, with enhancer activities defined by H3K4me1, H3K4me3 and H3K27ac signals, and promoter activities defined by its expression value. The second feature was similar to the first one, but the enhancer activity score was replaced by the expression levels of TFs that bind the enhancer. The third feature was the co-conservation of the enhancer and promoter sequences and the conservation of the synteny block across multiple species. The last feature was the genomic distance between the enhancer and promoter.

The trained model was applied to 12 human cell types by first identifying active enhancers in each cell type followed by extracting all promoters within a 2Mb window centered on the enhancer as their candidate targets. At a false discovery rate of 0.01, the resulting model predicted more than 440,000 unique enhancer-promoter associations in the 12 cell types in total. To validate the predictions, chromosome conformation capture coupled with quantitative PCR (3C-qPCR) was performed for 16 predictions and 13 of them could be validated. The predictions were also compared with interactions obtained from Hi-C and ChIA-PET, and reported eQTL-gene pairs. The results showed that IM-PET performed the best as compared to four other methods, namely nearest promoter, Ernst et al. (Ernst & Kellis 2010), Thurman et al. (Thurman et al. 2012) and PreSTIGE (Corradin et al. 2014).

The four features used in this work appear reasonable and biologically meaningful. The careful selection criteria for the training sets probably contributed to the good prediction results. Nevertheless, it should be noted that all the four features were not context-specific, including the activity correlation feature since only a single correlation was produced from each pair, as discussed above. Therefore, the method was unable to identify enhancer-target associations that are specific to particular contexts.

Databases useful for enhancer and enhancer-promoter association prediction

After discussing the features and latest methods used in computational prediction of enhancers and enhancer targets, here we list in Table 4 some of the popular databases that contain computationally predicted or experimentally validated enhancers and enhancer targets.

Table 4. Some databases that contain predicted or experimentally tested enhancers and enhancer targets

Database	Species	Description
dbSUPER (Hnisz et al. 2013)	Human and mouse	The first database of super-enhancers, containing a catalog of 66033 super-enhancers in 96 human and 5 mouse tissue/cell types.

		Provides a browser for functional analyses.
EI (Pennacchio et al. 2007)	Human and mouse	A database containing computational predicted tissue-specific enhancers based on TFBSs.
FANTOM5 Transcribed Enhancer Atlas (Andersson et al. 2014)	Human	An atlas of predicted enhancers based on eRNA; Contains 43011 computational predicted enhancers in total; Contains cell/organ/tissue-specific computational predicted enhancers; Contains computational predicted enhancer-promoter associations.
PEDB (Kumaki et al. 2008)	Human and mouse	A database of computational predicted enhancers based on conserved non-coding regions, TSSs and TFBSs.
PReMod (Ferretti et al. 2007)	Human and mouse	A computationally predicted CRM database based on TFBSs.
REDfly (Gallo et al. 2011)	<i>Drosophila</i>	A curated collection of known <i>Drosophila</i> CRMs and TFBSs; Contains enhancers <i>in vivo</i> ; Contains enhancer-promoter associations <i>in vivo</i> .
VISTA (Visel et al. 2007)	Human and mouse	Tested predicted human enhancers in mouse; Contains information on 2192 <i>in vivo</i> tested elements; 1154 elements with enhancer activity. (As of 4/15/2015)
ZEnBase (Navratilova et al. 2009)	<i>Zebrafish</i>	A database containing computational predicted enhancers based on conservation information.

Conclusions and discussions

Data processing

A fundamentally important but usually neglected topic in both enhancer prediction and enhancer target prediction is data processing. As with many problems in bioinformatics, different data processing strategies could result in huge differences in the results. For instance, Andersson et al. (Andersson et al. 2014) computed enhancer eRNA directionality scores based on the normalized CAGE data across 808 samples, which were normalized by converting tag counts to tags per million mapped reads (TPM) followed by normalization by

relative log expression (RLE) between samples. Our own analysis of this dataset shows that if a different normalization strategy is used, the resulting set of enhancers could become very different. In enhancer target prediction, whether taking log on gene expression levels could have big effects, especially when engaging a Pearson-correlation based measurement. Unfortunately, there is not a gold-standard normalization method that works best in all cases. Simple statistical analyses and plots of the data would help in the selection of the proper normalization method.

Feature usage

Good features play crucial roles in the prediction performance of machine learning methods, which we have discussed comprehensively above. Here, we discuss three important aspects of feature usage in enhancer and enhancer target predictions. First, as context specificity is an intrinsic characteristic of both enhancer activities and enhancer-target associations, we stress the importance of including context specific features. In the history of enhancer prediction, motifs and conservation were first used. These are “static” features, which means we could only use these features to judge whether a genomic region is an enhancer in some contexts, but not when (e.g., which developmental stage) and where (e.g., which cell type, cellular process) it would become active. Later, thanks to the boom of ChIP-seq data in a wide range of cell types, context-specific features such as HM and TF binding signals made it possible to perform cell-type specific enhancer predictions. In contrast, most current methods for enhancer target prediction use only static features. If the active enhancer-target associations in a given context are to be identified, one common strategy is to consider only the pairs involving an enhancer predicted/proved to be active in the context. Due to changes of chromosome conformation or other reasons, it is possible that an enhancer active in two different contexts regulates different genes in the two contexts. Novel methods that can utilize more context-specific information in directly predicting enhancer targets in a given context are called for.

A second interesting aspect is the relative importance of different features. When investigating a context with insufficient experimental data, and one is to perform additional experiments to get data for predicting enhancers or enhancer targets, it would be desirable to know what experiments are most cost-effective. Rajagopal et al. (Rajagopal et al. 2013) found a set of HMs that resulted in the best prediction accuracy, which partially answered this question. More generally, the relative importance of different types of features such as HM, TF binding, eRNA and DNA methylation is yet to be studied.

Another aspect is that a feature could be used for defining positive/negative examples, constructing the prediction model, or evaluating the performance of a model. For instance, P300 binding has been used in a number of studies for defining positive examples; Some studies use the enrichment of P300 binding signals as a way to partially validate the predictions; P300 could as well be used as a feature for building a model for predicting enhancer. One major current challenge is that given the limited number of features, one needs to determine which of them should be used in each of these three tasks, so that prediction

accuracy can be maximized while there is no “leakage” of information in the prediction process, i.e., having some information used both in training and validating a model. This problem is expected to be mitigated as more experimentally validated enhancers and enhancer targets become available.

Prediction validation

Prediction validation is a crucial part of every prediction task in bioinformatics. However, among the studies discussed, only a very small portion of the predictions made were tested experimentally. Obviously it is difficult to validate all predictions using highly accurate, low-throughput experimental assays due to the prohibitive cost. Another type of validations commonly performed is cross-checking the predictions with previously published experimental results such as ChIA-PET, Hi-C, 5C and eQTL-gene pairs for enhancer-target associations. One potential problem is that the predictions could be made in a context different from the one from which these public data were produced. Noise in these experimental data could be another issue. Also, some of these data only provide supporting evidence, but cannot completely prove the correctness of a prediction. For instance, a predicted enhancer-promoter association with *in vitro* ChIA-PET data support does not necessarily mean the enhancer-promoter interaction must have a regulatory role; It does not even guarantee the enhancer and the promoter are in contact *in vivo*. Having these shortcomings notwithstanding, including independent experimental supports would definitely help in evaluating and improving existing computational prediction methods. High-throughput assays such as STARR-seq, which has higher data variability but lower relative cost than low-throughput assays, could be a good choice for large-scale validations of computational predictions.

Training set design in supervised methods

The careful selection of training examples is key to the success of machine learning methods. In many bioinformatics problems, the design of a suitable negative training set is far from trivial. For instance, in enhancer prediction, the negative examples cannot be simply defined as randomly-selected regions not known to be enhancers, for these examples are too different from the positive examples in many aspects, and any model that distinguishes active regions in the genome from the inactive ones would probably separate the positive and negative examples well. In other words, the resulting model may not be useful for predicting enhancers, but just general active genomic regions including gene bodies and other types of regulatory elements. The rule of thumb is that the negative examples should not be “too negative”, i.e., they should share as many features as the positive examples as possible, except for the ones very unique to the positive examples. Alternatively, including a mixture of different types of negative examples could make it more robust.

Multiple-to-multiple relationships

After reviewing the current methods for association prediction, we notice that there are no existing methods that explicitly handle multiple-to-multiple relationships between enhancers and promoters. Every enhancer-promoter pair was considered independently in all the surveyed methods. Though the mechanisms of enhancer targeting are not completely clear yet, previous studies have shown that multiple enhancers (called shadow enhancers) controlling the same promoter could ensure the robust expression of the corresponding genes (Meireles-Filho & Stark 2009; Perry et al. 2011). New computational methods are needed to study the significance of modeling the effects of multiple enhancers and/or targets simultaneously.

Future outlook

Overall, we predict that context specificity and multiple-to-multiple relationships would be two important aspects that should be incorporated in future enhancer and enhancer target predictions.

Among all the features considered for the two tasks, we think eRNA is a promising feature for both tasks for two reasons: First, CAGE experiment is mature and economical and thus can be applied to many samples; Second, both eRNA and promoter activity are quantified in the same way based on CAGE tags, making the corresponding data easily comparable.

Since there are experimentally validated enhancers and enhancer targets, but the numbers are small, semi-supervised prediction methods that make use of both labeled examples and properties of unlabeled points could be more suitable than purely unsupervised or purely supervised methods.

Active learning is another direction worth pursuing. The active learning setting aims at acquiring new examples that can benefit the overall learning process most. In enhancer and enhancer target predictions, ambiguous cases (such as enhancers with intermediate levels of H3K27ac) could be most informative in refining prediction models.

Finally, we hope to see more collaboration between computer scientists and biologists in studying enhancers and enhancer targets, since the validation process is of utmost importance for evaluating the computational methods and providing insights for improving the methods.

References

- Ahituv, N. et al., 2005. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Human Molecular Genetics*, 14(20):3057–3063.
- Andersson, R. et al., 2014. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.
- Aran, D., Sabato, S. & Hellman, A., 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology*, 14(3):R21.
- Arnold, C.D. et al., 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077.
- Bailey, T.L. & MacHanick, P., 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):e128.
- Barski, A. et al., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837.
- Berman, B.P. et al., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biology*, 5(9):R61.
- Blow, M.J. et al., 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nature Genetics*, 42(9):806–810.
- Bonn, S. et al., 2012. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156.
- Boyle, A.P. et al., 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322.
- Carroll, S.B., 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, 134(1):25–36.
- Corradin, O. et al., 2014. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research*, 24(1):1–13.
- D'Alessandro, M. et al., 2003. Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: a report of four patients. *IEEE Transactions in Biomedical Engineering*, 50(5):603-615.
- Dawson, M.A. & Kouzarides, T., 2012. Cancer epigenetics: From mechanism to therapy. *Cell*, 150(1):12–27.

- Dekker, J. et al., 2002. Capturing chromosome conformation. *Science*, 295(5558):1306–1311.
- DeMare, L.E. et al., 2013. The genomic landscape of cohesin-Associated chromatin interactions. *Genome Research*, 23(8):1224–1234.
- Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.
- Dostie, J. et al., 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309.
- Duan, Z. et al., 2010. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367.
- ENCODE Project Consortium et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.
- Ernst, J. et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.
- Ernst, J. & Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216.
- Ernst, J. & Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825.
- FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.
- Ferretti, V. et al., 2007. PReMod: A database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Research*, 35(SUPPL. 1):D122-D126.
- Firpi, H.A., Ucar, D. & Tan, K., 2010. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, 26(13):1579–1586.
- Fu, Y. et al., 2014. FunSeq2 : a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10):480.
- Fullwood, M.J. et al., 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64.
- Gallo, S.M. et al., 2011. REDfly v3.0: Toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Research*, 39(SUPPL. 1):D118-D123.

- Giresi, P.G. et al., 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–885.
- He, B. et al., 2014. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21):E2191–E2199.
- Heidari, N. et al., 2014. Genome-wide map of regulatory interactions in the human genome. *Genome Research*, 24(12):1905–1917.
- Heintzman, N.D. et al., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318.
- Hnisz, D. et al., 2013. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947.
- Hoffman, M.M. et al., 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–841.
- Hoffman, M.M. et al., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476.
- Jin, F. et al., 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294.
- Kikuta, H. et al., 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, 17(5):545–555.
- Kim, T.-K. et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.
- Koller, D. & Friedman, N., 2009. *Probabilistic Graphical Models: Principles and Techniques*, MIT Press.
- Kumaki, Y. et al., 2008. Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14946–14951.
- Kwasnieski, J.C. et al., 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47):19498–503.
- Kwasnieski, J.C. et al., 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, 24(10):1595–1602.

- Larkin, D.M. et al., 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Research*, 19(5):770–777.
- Li, G. et al., 2010. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology*, 11(2):R22.
- Li, X.Y. et al., 2008. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biology*, 6(2):0365–0388.
- Lieberman-Aiden, E. et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.
- May, D. et al., 2011. Large-scale discovery of enhancers from human heart tissue. *Nature Genetics*, 44(1):89–93.
- Meireles-Filho, A.C. & Stark, A., 2009. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Current Opinion in Genetics and Development*, 19(6):565–570.
- Melnikov, A. et al., 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3):271–277.
- Navratilova, P. et al., 2009. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Developmental Biology*, 327(2):526–540.
- Nora, E.P. et al., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385.
- Park, P.J., 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics*, 10(10):669–680.
- Patwardhan, R.P. et al., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, 30(3):265–270.
- Pennacchio, L.A. et al., 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.
- Pennacchio, L.A. et al., 2007. Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17(2):201–211.
- Perry, M.W., Boettiger, A.N. & Levine, M., 2011. Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 108(33):13570–13575.

- Rada-Iglesias, A. et al., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.
- Rajagopal, N. et al., 2013. RFECs: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Computational Biology*, 9(3):e1002968.
- Ramos, Y.F.M. et al., 2010. Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Research*, 38(16):5396–5408.
- Rao, S.S.P. et al., 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680.
- Rhee, H.S. & Pugh, B.F., 2012. ChiP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, 21:21.24.
- Rhee, H.S. & Pugh, B.F., 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.
- Schug, J. et al., 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology*, 6(4):R33.
- Sharon, E. et al., 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530.
- Shlyueva, D., Stampfel, G. & Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews. Genetics*, 15(4):272–86.
- Su, J., Teichmann, S.A. & Down, T.A., 2010. Assessing computational methods of cis-regulatory module prediction. *PLoS Computational Biology*, 6(12):e1001020.
- Thurman, R.E. et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.
- Vaquerizas, J.M. et al., 2009. A census of human transcription factors: function, expression and evolution. *Nature Reviews. Genetics*, 10(4):252–263.
- Visel, A. et al., 2007. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research*, 35(SUPPL. 1):D88-D92.
- Visel, A., Rubin, E.M. & Pennacchio, L.A., 2009. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205.
- Wang, Z. et al., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903.

- Whyte, W.A. et al., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319.
- Won, K.-J. et al., 2008. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, 9:547.
- Xu, Y. et al., 2011. Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Molecular Cell*, 42(4):451–464.
- Yip, K.Y. et al., 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13(9):R48.
- Yip, K.Y., Cheng, C. & Gerstein, M., 2013. Machine learning and genome annotation: a match meant to be? *Genome Biology*, 14(5):205.
- Zhao, Z. et al., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11):1341–1347.

Keywords for Index

Enhancers, cis-regulatory elements, enhancer-target association, computational prediction, machine learning.