Predicting enhancers using a small subset of high confidence examples and co-training

Matthew R. Huska[*,1], **Anna Ramisch**[*,1], **Martin Vingron**[1], **and Annalisa Marsico**[2]

[1]**MPI for Molecular Genetics, Computational Molecular Biology, Berlin, Germany**
[2]**MPI for Molecular Genetics, RNA Bioinformatics, Berlin, Germany**
[*]**These authors contributed equally to this work**

ABSTRACT

Enhancers are important regulatory regions located throughout the genome, primarily in non-coding regions. Several experimental methods have been developed over the last several years to identify their location, but the search space is large and the overlap between the putative enhancer identified using these methods tends to be very small. Computational methods for enhancer prediction often use one large set of experimentally identified enhancer regions as input, and therefore rely critically on their correctness. We chose to take a different approach, and start with a high confidence set of 21 enhancer that are in the intersection of enhancers identified using three completely unrelated experimental approaches: deepCAGE, HiCap and classical enhancer reporter assays. Because this starting set is so small, we use a semi-supervised approach called co-training rather than a fully supervised approach to progressively predict enhancers from unlabeled regions. Using this approach we are able to outperform supervised learning as well as simpler semi-supervised learning methods and achieve an average area under the ROC curve of 0.84.

Keywords: enhancer prediction, co-training, semi-supervised learning

INTRODUCTION

Enhancers are genomic regions that function as cis-regulatory elements and have a key role in tissue- or condition-specific regulation of eukaryotic gene expression. They are typically short (less than 2 kilobases), are able to act over large distances, and are non-trivial to locate in the genome. The importance of distal regulatory elements like enhancers in human disease was suggested years ago by the observation in many genome-wide association studies that causal variants are very often identified distant from transcribed genes (Helgadottir et al., 2007). This suggestion has been confirmed in more recent studies showing that the disruption of the function of these regulatory elements can lead to changes in gene expression and disease phenotypes (Weedon et al., 2014; Lupiáñez et al., 2015).

Given the importance of enhancers in the context of gene regulation and disease, it is critical to be able to identify them in the genome. A multitude of methods have been developed over the last several years with the goal of genome-wide enhancer identification. Originally, cross-species DNA sequence conservation was used to identify enhancers, with the thought that important regulatory elements would be conserved through evolution (Pennacchio and Rubin, 2001; Nobrega, 2003; Pennacchio et al., 2006; Visel et al., 2007a, 2008). However, experimental validation of highly conserved regions using either in vivo or in vitro reporter assays showed that only about half of them were able to act as enhancers (Pennacchio et al., 2006; Visel et al., 2008). Later, methods were developed that could identify DNA/protein interactions on a genome-wide scale, including chromatin immunoprecipitation followed by microarray (ChIP-chip) and chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). These methods were used to try to identify enhancers based on the relationship between certain modified histones (mainly H3K4me1) or histone modifying proteins (such as p300).

More recently, methods for the large-scale identification of DNA-DNA interactions have been used for enhancer prediction as well (Sahlén et al., 2015; Chepelev et al., 2012; Li et al., 2012; Rao et al., 2014). This is based upon experimental evidence that enhancers function by physically looping to interact with the regions that they regulate (Müeller-Storm et al., 1989). Another method, STARR-seq (Arnold et al., 2013), in essence performs a large-scale version of a classic enhancer reporter assay that can measure ectopic enhancer activity, has been able to identify genomic regions

with enhancer potential genome-wide in drosophila, and in selected regions in humans. Lastly, enhancers were recently found to initiate RNA polymerase II (RNAPII) transcription, producing so-called eRNAs (Kim et al., 2010). Based on the FANTOM5 CAGE data from hundreds of cell lines and tissues, Andersson et al. (2014) identify more than 40,000 enhancer regions, together with their activation levels across human tissues, marked by the presence of bidirectional capped transcripts.

In addition to the experimental methods themselves, computational methods have been developed that use data from some of the preceding experimental methods as input. They can be divided into two broad classes of approaches: unsupervised methods, for example the genome segmentation algorithms Segway, ChromHMM and EpicSeg (Hoffman et al., 2012; Ernst and Kellis, 2012; Mammana and Chung, 2015), and supervised enhancer prediction methods such as RFECS (Rajagopal et al., 2013), EnhancerFinder (Erwin et al., 2013), and an SVM-based method by Lee et al. (2011). Unsupervised methods do not rely on any knowledge about already identified enhancer regions, but extract patterns (for instance, of different chromatin states) directly from the data – an advantage when no experimentally validated information is available. While this sounds like an advantage, the downside is that we do have knowledge of certain regions in the genome to be actual enhancers, but unsupervised methods do not take advantage of this information. In contrast, supervised methods rely critically on the existence of a large high-confidence labeled training set of known enhancer and non-enhancer regions. Most supervised computational methods are based on a single type of experimental data, for instance one of the experimental methods mentioned above, and use it as their large labeled training set. This is not ideal because each method only tests for one of the properties of enhancers that is currently believed to be necessary for their function: HiCap tests for looping, but regions can form loops without being active. ChIP-seq tests for the presence of certain histone modifications that are thought to be correlated with enhancer activity, but it is still unknown if this relationship is causal and the mechanisms for this relationship are still unknown. STARR-seq tests to see if the region can drive expression in a reporter, but this activity is ectopic and outside of the enhancer's native environment. The bidirectional transcription that Andersson et al. (2014) use to identify putative enhancer regions could be a mark of active enhancers (Andersson et al., 2014; Li et al., 2016), or simply a mark of accessible chromatin (Young et al., 2016). Beside the fact that these methods tend to test for only one property of enhancer activity, they are also large scale methods which each have their own biases, whether that is sequencing, amplification or other technical biases. This means that the putative enhancers identified by any of the previously mentioned methods are likely to contain a number of false positives that make them a poor starting point for supervised learning.

In contrast, we believe that a region that is identified by several experimental methods is more likely to be a true active enhancer, because it has several of the properties of enhancer regions: the ability to loop to a promoter or other enhancer, the ability to boost the expression of nearby genes, a chromatin environment that is thought to be conducive to enhancer activity, etc. Therefore we decided to start our enhancer prediction with a set of regions that are identified in not one but three separate experimental methods: HiCap, CAGE and ectopic enhancer assays as found in the Vista Enhancer Atlas.

This approach of choosing such a stringent set of enhancers when training a learning algorithm also has its disadvantages, in this case the size of the overlap between the enhancers predicted by all three methods is only 21 enhancer regions. These are likely not enough examples for a fully supervised approach, so instead we choose a method which is a compromise between supervised learning and unsupervised learning: semi-supervised learning.

Semi-supervised learning is a class of algorithms that utilizes, in addition to labeled data, also unlabeled data for training a model. Co-training is an instance of semi-supervised learning, and a self-labeling technique. As such, it enlarges the labeled training data in an iterative procedure by assuming that its own predictions (on the unlabeled data) tend to be correct. Typically it is employed in scenarios where the available labeled set is very small and the amount of unlabeled examples is large. This is ideal in our situation because we have a large amount of unlabeled data, the entire rest of the genome. The original co-training paper from Blum and Mitchell (1998) provides the theoretical assumptions for the co-training algorithm. They consider the setting in which the description of each example can be portioned into two distinct views or feature sets. Based on these two distinct views, two classifiers are initialized using just few labeled examples. The goal is to improve the performance of each individual classifier by using both views together, i.e. by taking advantage of an augmented labeled set containing confident predictions from both views. Then at each round of co-training each classifier chooses a subset of examples per class to add to the labeled set. Each classifier rebuilds from the augmented labeled set and the process repeats.

Augmenting training data in this way has led to numerous successful applications, for instance

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

in text and website category classification (Blum and Mitchell, 1998), prediction of geographical location (Riloff and Jones, 1999), word sense disambiguation (Yarowsky, 1995) and named entity classification (Collins and Singer, 1999). A successful bioinformatics application uses co-training to improve disease phenotype prediction from genotype by using a second classifier to impute the phenotype of unlabeled patients based on a second class of information: clinical health records (Roqueiro et al., 2015). A review study has shown that, in general, algorithms that make use of an independent split of the features outperform algorithms that do not (Nigam and Ghani, 2000).

In this paper we apply co-training to the problem of enhancer prediction because it allows us to start from a very small but high-confidence training set, and because we can deal with sequence features and epigenetic features separately in order to improve prediction performance. For sequence features, we use dinucleotide frequencies of each genomic region. The epigenetic features are input-normalized ChIP-seq counts for different histone marks, transcription factors and histone modifying proteins. We focus on mouse embryonic stem cells (mESCs) because ample experimental data is available for these cells. We demonstrate that co-training on a set of only 21 high confidence enhancer regions is able to achieve an average area under the ROC curve of 0.84.

The following Methods Section will introduce the data sets and the precise features used. It will also present the co-training algorithm as well as the validation set-up. Part of the Results Section is devoted to selection of parameter settings by first quantifying the effect of different parameter choices. Then we compare the prediction accuracy of our algorithm to the one of other competitor approaches, demonstrating that co-training produces very good results. Special emphasis is put on the learning behavior of co-training, which clearly reflects how starting from a small initial training set the method can iteratively select an increasingly better set of positive examples, thus slowly bootstrapping itself into recognizing enhancers.

METHODS

Experimentally Determined Enhancers

In order to obtain our high confidence set of mouse enhancers that are active in embryonic stem cells with which to start training our model, we combined data from several sources which attempt to identify enhancers using different experimental methods. Putative enhancers in mouse were retrieved from the Vista Enhancer Browser database (Visel et al., 2007b), the FANTOM5 project's CAGE-based bidirectional transcription enhancer set (Andersson et al., 2014) and HiCap data from Sahlén et al. (2015).

HiCap enhancers

We downloaded putative enhancers in mESCs from the table of high-resolution, genome-wide map of promoter-enhancer and enhancer-enhancer interactions determined with the HiCap technique (Sahlén et al., 2015). In detail, HiCap allows the identification of 3D chromatin interactions anchored on gene promoters by using a combination of proximity-based ligation procedures, as in the Hi-C method (Lieberman-Aiden et al., 2009), together with sequence capture of annotated promoters. Distal regions connected to promoters or other distal regions, where the interaction is supported with three or more reads in both replicates, were defined by the authors as "putative enhancers" and used in our analysis. Our final set of putative HiCap enhancers comprises 71,698 unique genomic regions, which can be involved in more than one interaction.

FANTOM5 enhancers

From the FANTOM5 Transcribed Enhancer Atlas (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/) we downloaded 44,150 putative enhancer regions. These enhancers were defined by detecting bidirectional transcription at sites of enriched deepCAGE signal in various mouse cell lines and tissues (Andersson et al., 2014). The entire mouse permissive enhancers "phase 1 and 2" set was used.

VISTA enhancers

We downloaded 323 putative enhancer regions in embryonic mouse tissues from the VISTA Enhancer Browser (Visel et al., 2007b). These regions, initially identified by extreme evolutionary sequence conservation or by ChIP-seq, were experimentally validated in a transgenic mouse assay.

Partitioning of the data set

The entire data set was partitioned into three distinct sets: a high confidence labeled set L, a hold-out validation set and an unlabeled set U. Each genomic region can only belong to one of these sets. The

labeled set L and the unlabeled set U change their composition during the iterative procedure in the co-training algorithm.

Initial labeled set

The initial training set L is an almost fully balanced set of 21 positive and 28 negative examples. More precisely, the positive set is the intersection of enhancers defined by VISTA, FANTOM5 and HiCap experiments (as described above), and as such represents high-confidence enhancer regions. The negative set is composed of 14 promoter regions and 14 intergenic regions of size 300 bps, which were randomly chosen from the rest of the genome.

Validation set

Our validation set consists of 500 positive and 500 negative examples. We chose positive examples for our validation set randomly from intersections of only two of the three putative enhancer sets (see Figure 1). The negative set is built from 250 randomly chosen promoter regions and 250 randomly chosen regions of size 300 bps from the rest of the genome.

Initial unlabeled set

The unlabeled set U consists of 111,183 identified putative enhancers from VISTA, FANTOM5 and HiCap experiments, 21,359 promoters and 99,736 randomly chosen intergenic regions.

Predictive Features

ChIP-seq data

The following mESC ChIP-seq data sets were used for building the first active enhancer classifier: H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3, H3K36me3, H2AZ and corresponding input (GEO series GSE36114, E14 Day0 samples); and Pol2, p300, CTCF and corresponding input (GEO series GSE29184, samples GSM723019, GSM723018, GSM723015 and GSM723020). The data were pre-processed as described in Juan et al. (2016). Briefly, sra files were transformed into fastq files with sra-toolkit (v2.1.12) and aligned to the reference mm9/NCBI37 genome with BWA v0.5.9-r16 (Li and Durbin, 2009) allowing zero to one mismatches. We counted the overlapping ChIP-seq reads in the genomic regions of interest, namely enhancers, promoters and intergenic regions using bamsignals v1.2.1 (Mammana and Helmuth, 2014). Read counts in genomic regions where divided by the input read counts and log normalized.

Sequences

We used the mouse genome as provided by UCSC (mm9, Jul. 2007) and computed the dinucleotide frequencies of our regions of interest using the oligonucleotideFrequency function from the Biostrings v2.38.4 package in R v3.2.3.

Co-training

The co-training algorithm explicitly uses a feature set split when learning from labeled and unlabeled data. Its approach is to initialize classifiers based on distinct feature sets using just the few labeled high-confidence examples. In an iterative procedure the labeled data set is enlarged by assuming that the predictions of the classifiers tend to be correct.

We applied co-training to the problem of classifying active enhancers versus random intergenic and promoter regions. Our feature space $X = X_1 \times X_2$ is composed of two different views of an example region $x = (x_1, x_2)$. That is, for each genomic region x we considered x_1 to be a feature vector corresponding to the ChIP-seq data described above, and x_2 to be the vector of dinucleotide frequencies. Let g_1 and g_2 denote logistic regression classifiers based on view X_1 and X_2, respectively. From now on we will refer to g_1 as the ChIP-seq-based and to g_2 as the sequence-based classifier. As a first step, g_1 and g_2 are trained on the initial training set L. The classifiers' most confident predictions on unlabeled regions are used to enlarge the training set iteratively. More specifically, we sample two subsets U_1 and U_2 of size $n_u \leq |U|$ from the unlabeled set, and predict the class probabilities for the contained regions using g_1 and g_2, respectively. Then we choose the most confident regions from both sets U_1 and U_2 according to confidence criterion c_{conf}. Regions with a class probability greater than 0.5 are added to the labeled set L as positive examples, regions with probability less than 0.5 as negative examples. Contradictively labeled regions in the intersection of U_1 and U_2, i.e. regions x for which $g_1(x_1) \neq g_2(x_2)$, are not added to L, but remain in the unlabeled set U. The co-training algorithm used in this paper is a re-adaptation of the algorithm from Blum and Mitchell and it is schematically described in Algorithm 1. As an output we get the two classifiers g_1 and g_2 trained on an augmented labeled set L. Based on g_1 and g_2 we build an overall classifier g with which we predict class labels for new genomic regions in the validation set. This is done by multiplying together the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

class probabilities predicted with g_1 and g_2, and then re-normalizing these class probability scores so they sum up to one.

Algorithm 1 Co-training

Input: high-confidence labeled set L, unlabeled set U, parameter vector $\theta = (n_u, c_{conf})$
Output: classifiers g_1 and g_2 trained on augmented labeled set L

while stopping criterion is not met **do**
  for $i = 1, 2$ **do**
    train classifier g_i on L
    sample subset U_i of size n_u from U
    predict class probabilities of regions in U_i with g_i
    determine most confidently predicted regions in U_i based on criterion c_{conf}
    label most confidently predicted regions in U_i
  end for
  if intersection of U_1 and U_2 is not empty ($U_1 \cap U_2 \neq \emptyset$) **then**
    discard contradictively labeled regions ($g_1(x_1) \neq g_2(x_2)$)
  end if
  add labeled regions from U_1 and U_2 to L
end while

RESULTS AND DISCUSSION

The enhancer prediction problem is mostly addressed by supervised learning algorithms based on chromatin features and/or sequence properties that need a large quantity of labeled examples to perform well. Such approaches rely on genome-wide experimental data to define enhancers, making the labeled training set highly dependent on the specific characteristics of the experiment or high-throughput technology that is used. In this work we first inspected the overlap of genomic coordinates between FANTOM5 CAGE-based enhancers, HiCap 3D chromatin interaction data and experimentally validated enhancers from the VISTA Enhancer Browser. The three data sets contain 44,150, 71,698 and 323 putative enhancers respectively, but their intersection is limited to 21 as can be seen in Figure 1. Such a small intersection indicates that it is a non-trivial and very challenging task to design a training set for a supervised enhancer classification task. One possible choice would be to use one of the three putative enhancer sets to train the model. In this scenario, the training set is large and fully supervised learning methods are expected to have a good performance, but it also means adding more noise and technique-specific bias. Another option is to choose the small subset of 21 enhancers at the intersection as a positive set. Given that enhancers at the intersection are supported by three completely different experimental techniques, we opted for the latter choice, and used a semi-supervised, rather than a fully supervised approach. In particular, we implemented the algorithm from Blum and Mitchell, where the description of each example is partitioned into two distinct views. Our ultimate goal is to see whether in this multi-view feature setting, unlabeled examples can help to predict active enhancers, starting from very few high-confidence labels and progressively adding unlabeled examples. Since the intersection of the three sources of putative enhancers is so small, it is not possible to use a subset of these very confident regions for validation. In order to decrease the probability of false positive enhancers as much as possible, we sampled positive regions from the pairwise intersections of the three putative enhancer sources for our validation set.

Experiments

Experiments were conducted to determine whether our co-training approach could successfully use unlabeled data to predict active enhancers in mouse ES cells and outperform both standard supervised learning methods and another semi-supervised method called self-training, which does not make use of a feature split. In more detail, we compared co-training to the following methods and feature sets:

- semi-supervised self-training based on logistic regression with
  - ChIP-seq-based features
  - sequence-based features
  - All features (ChIP-seq and sequence-based)
- supervised logistic regression with
  - ChIP-seq-based features
  - sequence-based features
  - All features (ChIP-seq and sequence-based)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

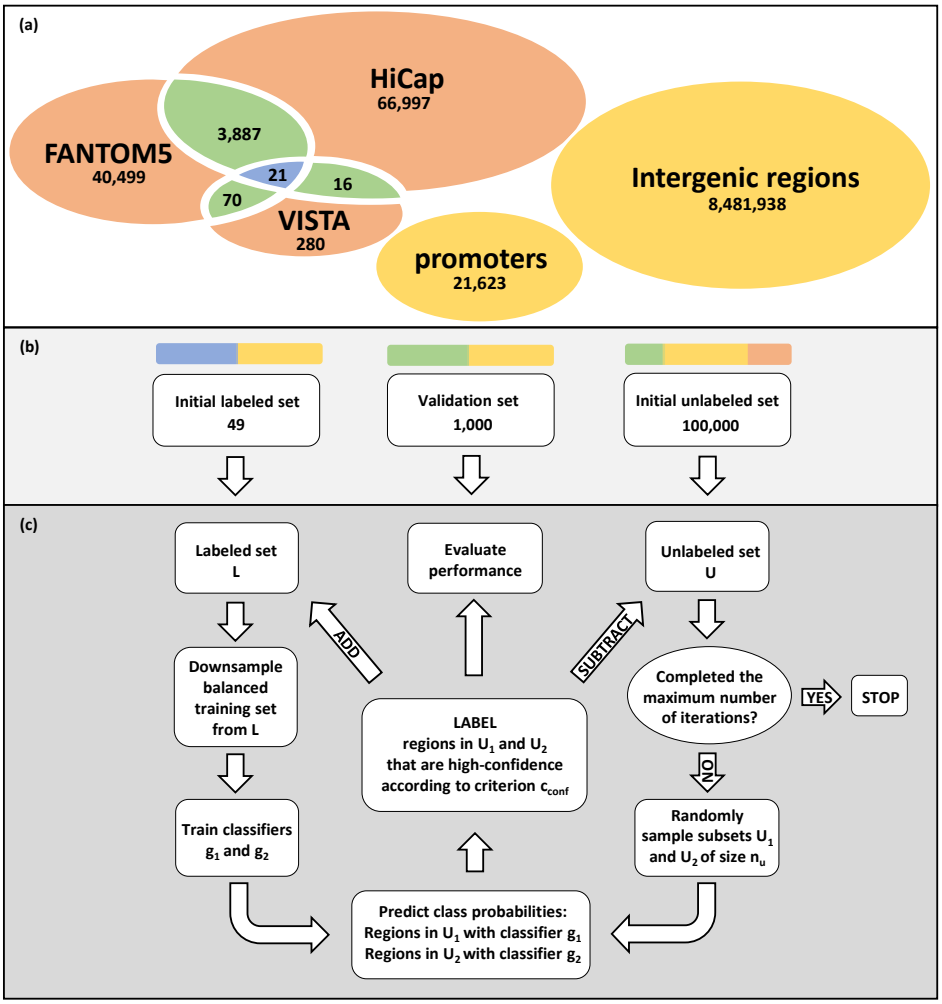

Figure 1. (a) Venn diagram of putative enhancer sets, promoters and intergenic regions in the mouse genome. The intersection of all the putative enhancer sets includes 21 "high-confidence" regions. (b) Size and composition of the validation set, and the initial labeled and unlabeled sets. The colors in the bars refer to the genomic regions contained in the sets. (c) Workflow of our co-training method.

Parameter tuning

Both the co-training and self-training algorithms have several important parameters that need to be chosen. The parameters we focused on were (i) the size n_u of the unlabeled sets U_1 and U_2 that are sampled at each iteration step, and (ii) the criterion c_{conf} for defining the most confidently labeled examples that are added to the training set after each iteration.

In order to select good parameters for each method, we tuned these parameters using a grid search and selected the parameter combination that resulted in the highest average AUC ROC across four randomly chosen validation sets, with four random seeds per validation set, resulting in 16 runs of each parameter set total. The parameters that were tested were $n_u = 100, 200, 500, 5000$, and two criteria for selecting the most confidently labeled examples were evaluated. The first criterion was to just take a fixed number of examples per iteration, 6, 10, 25, 50, or 100 examples, and add them to the labeled set L. An additional parameter for this criterion was whether the selected examples should be forced to be equally split between positive and negative examples, or if the highest confidence examples are selected regardless of which class they belong to. The second criterion for choosing the most confidently labeled examples was using a score cutoff. Logistic regression gives scores between 0 and 1, so we evaluated using a score cutoff $s = 0.05$, 0.1, or 0.25, and all regions with a predicted probability $p \leq s$ or $p \geq 1 - s$ were added to the labeled set L. The optimal parameters for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

each method and the resulting AUC ROC is show in Table 1.

Table 1. Optimal Parameters For Each Method

Method and Features	n_u	c_{conf}	mean AUC
Co-training	100	top 50 (force balanced)	0.84
Self-training All	5000	top 6	0.71
Self-training ChIP-seq	500	top 6	0.80
Self-training Sequence	5000	top 6 (force balanced)	0.68

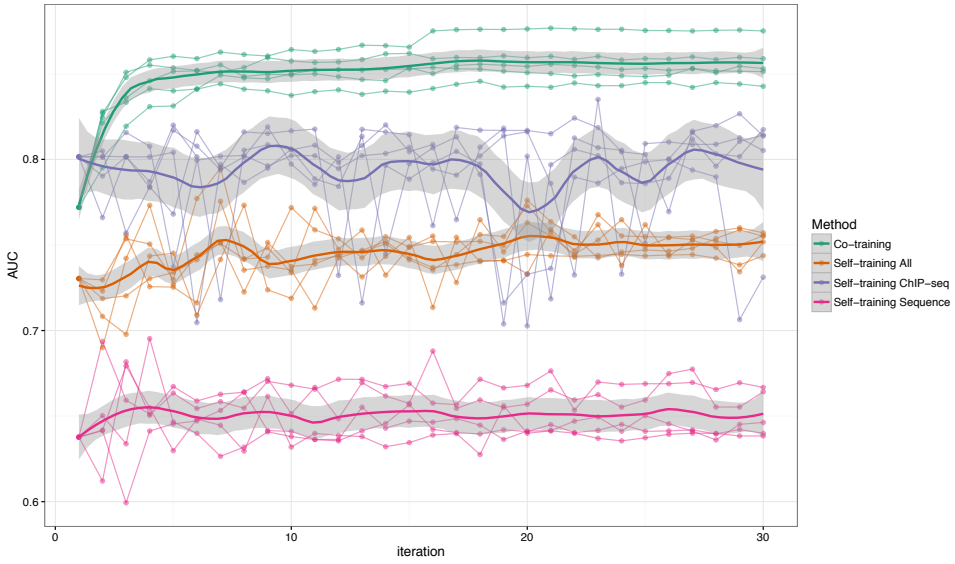

Figure 2. The performance of each classifier on one of the held out validation sets after each iteration of training. The performance measure used is the area under the receiver operating characteristic curve. Each method was run five times with a different random seed each time, which affects the random sample of unlabeled examples that is considered at each iteration. LOESS regression was used to plot a smoothed estimate for each method and the 95% confidence interval around these estimates is shown in grey.

Performance of co-training compared to other settings

Using the best set of parameters for each method as identified using the procedure outlined in the previous section we monitored the performance of each method at each iteration on 20 different validation sets (for one example, see Figure 2). In every case, and independently from the choice of the validation set, our co-training method performs on average better than the other methods (see Figure 3). In addition, unlike the other methods the classification performance (AUC ROC) reliably improves after incorporating unlabeled examples into the prediction. This seems to be due to the situation that even with very few features (10 ChIP-seq experiments and/or dinculeotide frequencies), most methods are prone to over-fitting when starting with such a small training set. This can be observed by the fact that the self-training methods do not reliably improve their performance as they iteratively add new examples to their labeled sets. Co-training helps us avoid this over-fitting by mixing the best predictions of the two separate views and incorporating these predictions into the following iterations that would not be selected by each classifier alone.

In the future, other feature sets containing higher information content than dinucleotide frequencies, for example transcription factor binding site motif match scores, could be used in this setting to further improve the classifier performance as well as biological interpretability. Likewise, more suitable choices of both the initial labeled set and validation set could be made, especially as more experimental data becomes available. Finally, it is important to see how this method performs in other

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

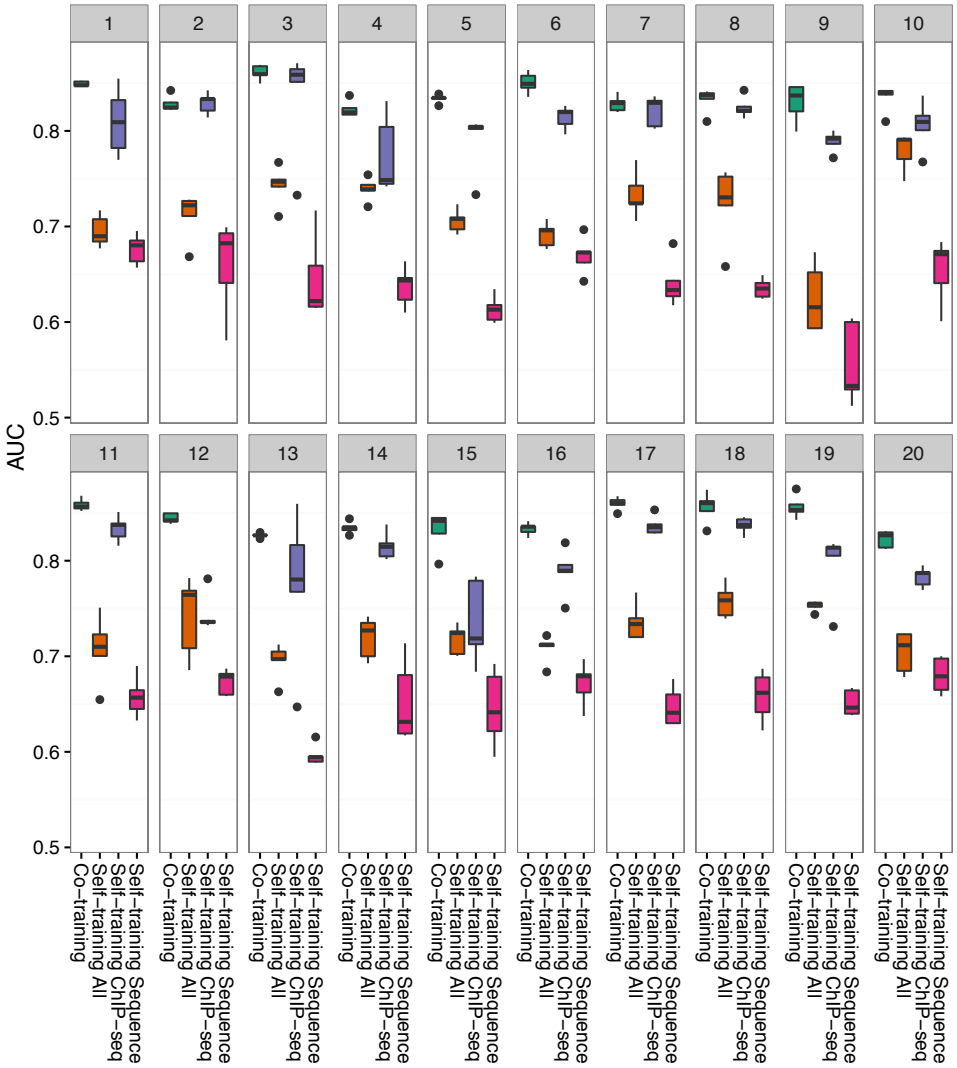

Figure 3. To investigate how sensitive the results are to the choice of validation set, we ran all methods on 20 different validation sets each comprised of a randomly selected 500 positive regions and 500 negative regions made up of 250 promoters and 250 other genomic regions. For each validation set, each method was run 5 times using a different random seed.

cell types in order to confirm that the results observed here generalize to other cell lines, tissues and organisms.

CONCLUSION

The selection of a large training set of proven enhancers is challenging due to possible biases associated to experimental techniques for determining enhancers, which makes it difficult to apply supervised learning algorithms to the problem of enhancer prediction. We therefore chose a different approach, starting with a small high-confidence set of experimentally validated enhancers, we use a semi-supervised learning method called co-training which lets two classifiers cooperatively take advantage of information from a large unlabeled set of genomic regions. As a result of this process a small learning set suffices to train a highly accurate classifier for this difficult prediction problem. We show that the use of a small initial high-confidence training set, unlabeled data, and two separate feature views, has potential for significant benefits over other methods in the enhancer prediction task.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016

ACKNOWLEDGMENTS

Support by Deutsches Epigenomprojekt (DEEP) and SFB-TR 84 is gratefully acknowledged.

REFERENCES

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (New York, N.Y.)*, 42(3):255–9.

Blum, A. and Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100.

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell research*, 22(3):490–503.

Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216.

Erwin, G. D., Truty, R. M., Kostka, D., Pollard, K. S., and Capra, J. A. (2013). Integrating diverse datasets improves developmental enhancer prediction. 37232(615):33.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., Masson, G., Gudbjartsson, D. F., Magnusson, K. P., Andersen, K., Levey, A. I., Backman, V. M., Matthiasdottir, S., Jonsdottir, T., Palsson, S., Einarsdottir, H., Gunnarsdottir, S., Gylfason, A., Vaccarino, V., Hooper, W. C., Reilly, M. P., Granger, C. B., Austin, H., Rader, D. J., Shah, S. H., Quyyumi, A. A., Gulcher, J. R., Thorgeirsson, G., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316(5830):1491–1493.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–476.

Juan, D., Perner, J., de Santa Pau, E. C., Marsili, S., Ochoa, D., Chung, H.-R., Vingron, M., Rico, D., and Valencia, A. (2016). Epigenomic co-localization and co-evolution reveal a key role for 5hmc as a communication hub in the chromatin network of escs. *Cell reports*, 14(5):1246–1257.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome research*, 21(12):2167–80.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–1760.

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics*, 17(4):207–223.

Lieberman-Aiden, E., Berkum, N. L. V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., and Mirny, L. A. (2009). of the Human Genome. *Analyzer*, 33292(October):289–293.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

Mammana, A. and Chung, H.-R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):1–12.

Mammana, A. and Helmuth, J. (2014). bamsignals: Extract read count signals from bam files. R package version 1.2.0.

Müeller-Storm, H. P., Sogo, J. M., and Schaffner, W. (1989). An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell*, 58(4):767–77.

Nigam, K. and Ghani, R. (2000). Understanding the behavior of co-training. In *Proceedings of KDD-2000 Workshop on Text Mining*.

Nobrega, M. A. (2003). Scanning Human Gene Deserts for Long-Range Enhancers. *Science*, 302(5644):413–413.

Pennacchio, L. a., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. a., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

Pennacchio, L. a. and Rubin, E. M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nature reviews. Genetics*, 2(2):100–109.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Computational Biology*, 9(3):e1002968.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Riloff, E. and Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 474–479.

Roqueiro, D., Witteveen, M., Anttila, V., Terwindt, G., van den Maagdenberg, A., and Borgwardt, K. (2015). In silico phenotyping via co-training for improved phenotype prediction from genotype. *Bioinformatics*, 31(12):i303–i310.

Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., Albert, T. J., Lundeberg, J., and Sandberg, R. (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, 16(1):156.

Visel, A., Bristow, J., and Pennacchio, L. a. (2007a). Enhancer identification through comparative genomics. *Seminars in Cell & Developmental Biology*, 18(1):140–152.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. a. (2007b). VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic acids research*, 35(Database issue):D88–92.

Visel, A., Prabhakar, S., Akiyama, J. a., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., and Pennacchio, L. a. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature genetics*, 40(2):158–60.

Weedon, M. N., Cebola, I., Patch, A.-m., Flanagan, S. E., De Franco, E., Caswell, R., Rodríguez-seguí, S. a., Shaw-Smith, C., Cho, C. H.-H., Lango Allen, H., Houghton, J. a. L., Roth, C. L., Chen, R., Hussain, K., Marsh, P., Vallier, L., Murray, A., Ellard, S., Ferrer, J., Hattersley, A. T., Franco, E. D., Allen, H. L., Pancreatic, I., and Consortium, A. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature genetics*, 46(1):61–4.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

Young, R. S., Kumar, Y., Bickmore, W. A., and Taylor, M. S. (2016). Bidirectional transcription marks accessible chromatin and is not specific to enhancers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2407v1 | CC BY 4.0 Open Access | rec: 1 Sep 2016, publ: 1 Sep 2016