Variational Infinite Heterogeneous Mixture Model for Semi-supervised Clustering of Heart Enhancers

Tahmid F. Mehdi[1], Gurdeep Singh[2], Jennifer A. Mitchell[2,3] and Alan M. Moses[1,2,4,*]
[1] Department of Computer Science, University of Toronto, Toronto, M5S 2E4, Canada,
[2] Department of Cell & Systems Biology, University of Toronto, Toronto, M5S 3G5, Canada,
[3] Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, M5S 3G5, Canada
[4] Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, M5S 3B3, Canada
[*] To whom correspondence should be addressed.

Abstract

Motivation: Mammalian genomes can contain thousands of enhancers but only a subset are actively driving gene expression in a given cellular context. Integrated genomic datasets can be harnessed to predict active enhancers. One challenge in integration of large genomic datasets is the increasing heterogeneity: continuous, binary and discrete features may all be relevant. Coupled with the typically small numbers of training examples, semi-supervised approaches for heterogeneous data are needed; however, current enhancer prediction methods are not designed to handle heterogeneous data in the semi-supervised paradigm.

Results: We implemented a Dirichlet Process Heterogeneous Mixture model that infers Gaussian, Bernoulli and Poisson distributions over features. We derived a novel variational inference algorithm to handle semi-supervised learning tasks where certain observations are forced to cluster together. We applied this model to enhancer candidates in mouse heart tissues based on heterogeneous features. We constrained a small number of known active enhancers to appear in the same cluster, and 47 additional regions clustered with them. Many of these are located near heart-specific genes. The model also predicted 1176 active promoters, suggesting that it can discover new enhancers and promoters.

Availability: We created the 'dphmix' Python package: https://pypi.org/project/dphmix/

Contact: alan.moses@utoronto.ca

Introduction

Enhancers are cis-regulatory elements in DNA that can influence expression levels of target genes when bound to transcription factors (TFs). They are thought to exist in at least three states of differing activity: active, primed and poised enhancers (Calo and Wysocka, 2013), such that only a subset of bound regions (active enhancers) play a role in gene regulation in a given cellular context. Although different states are distinguished by differential patterns of histone modifications and transcriptional regulator recruitment, systematically classifying the state of an enhancer remains a challenge (Zentner et al., 2011; Catarino and Stark, 2018).

Modern genomic data is highly heterogeneous and may contain continuous (e.g. histone modification levels), binary (e.g. TF-binding) and discrete features (e.g. counts of methylated sites) and, in principle, supervised machine learning methods can be used to identify active enhancers with integrated heterogeneous genomics data. Enhancer activity, however, is highly tissue-specific (Bulger and Groudine, 2011), and there are few tissues for which large numbers of active enhancers have been identified, limiting the application of recent supervised enhancer prediction approaches such as DECRES (Li et al., 2018) and REPTILE (He et al., 2017) to identify active enhancers in tissues of interest. Furthermore, currently, supervised methods have not been designed to predict enhancer states other than active and inactive enhancers.

Clustering (or unsupervised) techniques could, in principle, identify clusters of genomic regions that are enriched with different states of enhancers without large training sets. Unsupervised methods like ChromHMM (Ernst and Kellis, 2012) and Segway (Hoffman et al., 2012) can predict enhancers; however, these methods were designed for genome segmentation rather than enhancer state prediction over enhancer candidates (like TF-bound regions). Ideally, the small numbers of experimentally validated active enhancers should be used if possible: this motivates the development of semi-supervised approaches that can integrate heterogeneous data.

To integrate high-dimensional genomic data and predict enhancer states over enhancer candidates, we developed a variational Dirichlet Process Heterogeneous Mixture (DPHM or an infinite heterogeneous mixture) model that infers Gaussian, Bernoulli and Poisson distributions over continuous, binary and non-negative discrete features, respectively. To take advantage of small labeled training sets where available, we derive a novel variational inference algorithm for a semi-supervised DPHM model that forces a subset of the data (like experimentally validated enhancers) to cluster together. Our Bayesian model also has the advantages that 1) the number of clusters, or enhancer states, is inferred from the data and 2) the number of hyperparameters does not grow with the number of clusters, simplifying inference for heterogeneous data integration. The DPHM model outperformed Gaussian mixture models in clustering synthetic heterogeneous datasets in unsupervised and semi-supervised settings. The DPHM model can also outperform k-means, given adequate initializations, even when k-means is given the correct number of clusters.

To illustrate the power of the DPHM model to integrate heterogeneous genomic data, we employed it to predict new enhancers based on heterogeneous features from the Encyclopedia of DNA Elements (ENCODE) Project (ENCODE Project Consortium, 2012). We applied the semi-supervised DPHM model to a dataset of 6,209 genomic regions bound by Nkx2-5, a master regulator in heart development, in embryonic mouse heart with the constraint that a set of known active enhancers have to cluster together. Through our novel variational inference algorithm, 47 new regions clustered with the experimentally known active enhancers in this tissue. Furthermore, we discovered 5 large classes of genomic regions in the data. Some classes, including the class with the known active enhancers, were significantly ($q < 0.05$) enriched with various biological processes. Another class, enriched for house-keeping genes, appears to contain active promoters. Moreover, each functional class of enhancers was enriched with at least 60 different TF-binding motifs and some motifs can be utilized to discriminate the classes from each other. Our analysis indicates that the semi-supervised DPHM model is a principled Bayesian method for discovering biologically relevant clusters in heterogeneous genomic data in the semi-supervised learning paradigm.

Methods

Data

Dupays et al. (2015) identified genomic regions bound by Nkx2-5 using chromatin immunoprecipitation sequencing (ChIP-seq) from embryonic mouse heart tissue. Genomic coordinates for the binding sites were converted from the GRCm37/mm9 to the GRCm38/mm10 genome build. To incorporate information about the functional conservation of the sites, we identified regions of the human genome (GRCh38/hg38 build) whose DNA sequences are alignable to the sites using LiftOver (Hinrichs et al., 2006). We found that 6209/7246 of the binding sites in the mouse genome had human orthologs. 109 of the regions overlapped with experimentally validated developmental enhancers in mouse heart tissue, identified by the VISTA Enhancer Browser (Visel et al., 2007). Features for the regions were generated from ENCODE datasets (ENCODE Project Consortium, 2012) for mouse and human heart tissues. Supplementary Table 1 lists all of the datasets that we used to extract features. We extracted 81 mouse features and 33 human features.

The mouse features were comprised of 68 continuous features, 5 binary features and 8 ordinal features. The continuous features were composed of 65 histone modification features and 3 DNase-seq features. For histone modification features, we used 65 different ChIP-seq datasets for different histone modifications and time points. For DNase-seq features, we used 3 different DNase-seq datasets for 3 different time points. The binary features were created from 5 different TF ChIP-seq datasets (2 datasets for CTCF binding at 2 different time points, 2 datasets for P300 at 2 different time points and 1 dataset for POL2 binding). Finally, ordinal features were generated from whole-genome bisulfite sequencing (WGBS) data, which measures DNA methylation (DNAm) levels of CpG sites. We included WGBS data because He et al. (2017) reported that DNAm data can help predict enhancers. We used 8 WGBS datasets

for 8 different time points. For each observation, continuous features were extracted by taking the maximum histone modification ChIP-seq or DNase-seq signal across its corresponding genomic region in mouse. Binary features were extracted, for each observation, to indicate whether the TF was bound to the corresponding region in mouse (1 if there is a ChIP-seq peak in the region, 0 otherwise). Finally, for each observation, ordinal features were extracted by counting the number of highly methylated (DNAm level > 0.5) CpG sites across the corresponding region in mouse.

Human features consisted of 32 continuous features and 1 ordinal feature. Continuous features were composed of 7 histone modification features and 25 DNase-seq features. For histone modification features, we used 7 different ChIP-seq datasets for different histone modifications, time points and samples. For DNase-seq features, we used 25 DNase-seq datasets for different time points and samples. The ordinal feature was created using WGBS data from an embryonic time point. We extracted human features using the same methods that we employed to extract mouse features.

Dirichlet Process Heterogeneous Mixtures

The DPHM model takes observations with heterogeneous features and clusters them based on similarities between their features. Continuous, binary and non-negative discrete features are assumed to follow Gaussian, Bernoulli and Poisson distributions, respectively, and are assumed to be mutually independent so the conditional likelihood of an observation is the product of distributions, with cluster-specific parameters, for each feature. In this section, we assume the data has r_g, r_b and r_p Gaussian, Bernoulli and Poisson features, respectively.

Figure 1 shows the latent variables of the semi-supervised DPHM model. μ and τ contain the means and precisions of the Gaussian features, respectively, p contains the probability parameters for the Bernoulli features and λ contains the average rate parameters for the Poisson features. The subscripts on the variables denote their feature indices and associated clusters. For example, μ_{tj} represents the mean parameter for the j-th Gaussian feature and cluster t.

Priors

Distribution parameters (μ, τ, p and λ) are drawn from their conjugate priors (Bishop, 2006). The conjugate priors for Gaussian, Bernoulli and Poisson distributions are NormalGamma (NG), Beta and Gamma distributions, respectively. Since features are assumed to be mutually independent, the joint conjugate prior for the distribution parameters can be expressed as:

$$\mu, \tau, p, \lambda \sim \prod_{j=1}^{r_g} NG(\nu_j, \rho_j, a_j, b_j) \prod_{k=1}^{r_b} Beta(\gamma_k, \delta_k) \prod_{l=1}^{r_p} Gamma(\varepsilon_l, \zeta_l) \quad (1)$$

where hyperparameters v_j , ρ_j , a_j and b_j control the prior mean, precision, shape and rate parameters, respectively, of the NG distribution that generates the mean and precision for the *j*-th Gaussian feature. γ_k and δ_k represent the prior shape parameters of the Beta distribution that creates the probability parameter for the *k*-th Bernoulli feature. ε_l and ζ_l are the prior shape and rate parameters, respectively, for the Gamma distribution that generates the average rate parameter for the *l*-th Poisson feature. The conjugate prior eases computations in the algorithm because the cluster-specific posterior distributions of μ, τ, p and λ will be members of the conjugate prior's family; we refer to this family as 'NGBG'.

Cluster assignments (denoted by *c*) for each observation are drawn from multinomial distributions, and their prior parameters are mixing weights for the clusters. Mixing weights are constructed through the truncated stick-breaking process (Ishwaran and James, 2001) that sets an upper bound, *T*, on the number of clusters. Mixing weights are completely determined by stick-breaking variables $\{v_t\}_{t=1}^{T}$ that depend on α, a hyperparameter that controls cluster sizes. Although there are *T* clusters, some of them will be empty if *T* is large enough and can be ignored in downstream analyses. Cluster assignments, stick-breaking variables and distribution parameters form the latent variable space, while α and parameters of the NGBG prior form the hyperparameter space of the DPHM model.

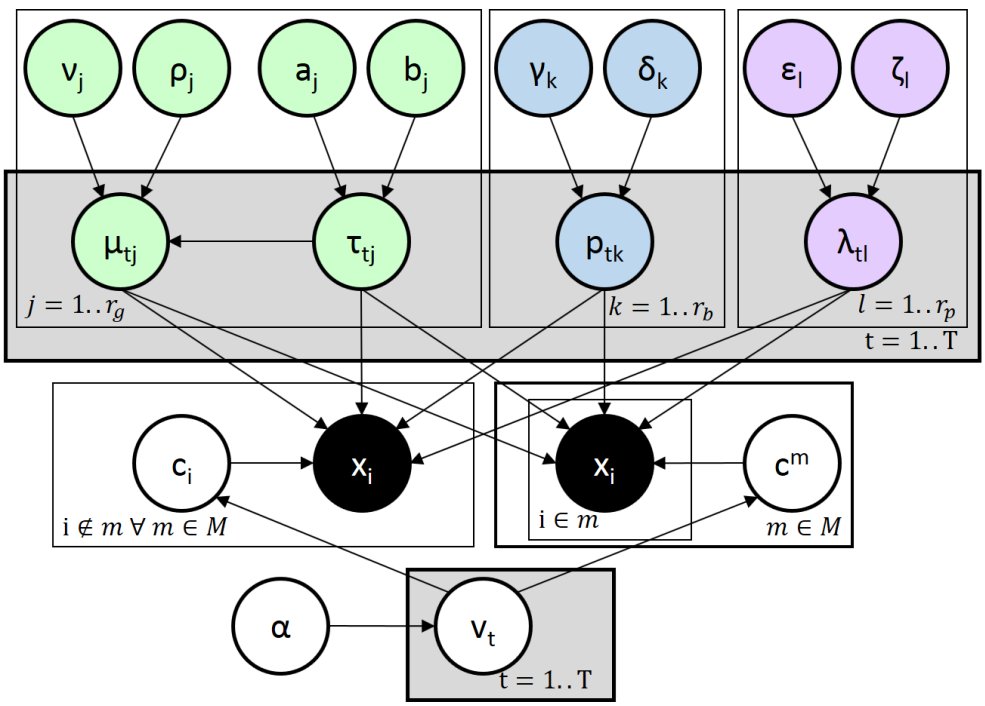

Fig. 1. A Bayesian network depicting the dependencies between hyperparameters, latent variables and observations in the DPHM. Each observation, x_i, depends on their cluster assignment (c_i for observations that are not in any must-link constraints and c^m for observations in a must-link constraint *m*) and distribution parameters (μ, τ, p and λ). Distribution parameters depend on parameters of the NGBG prior (v, ρ, a, b, γ, δ, ε and ζ). Distribution parameters and hyperparameters related to Gaussian, Bernoulli and Poisson features are green, blue and purple, respectively. Each cluster assignment depends on the v_ts that are generated through the stick-breaking construction. *M* represents a set of must-link constraints for semi-supervised clustering.

Variational Inference for Semi-supervised DPHM Models

We use variational inference (Beal, 2003; Blei and Jordan, 2006) to fit the DPHM model to a dataset *X*. Variational inference approximates the true posterior of the latent variables with a variational distribution *q* by maximizing the evidence lower bound (ELBO):

$$ELBO(q) = \mathbb{E}[\ln P(X, \mu, \tau, p, \lambda, c, v)] - \mathbb{E}[\ln q(\mu, \tau, p, \lambda, c, v)] \quad (2)$$

We optimize over the mean-field variational family so *q* factorizes into the product of variational densities for the latent variables. The variational density for a cluster assignment is parameterized by a 'cluster probability vector' that contains probabilities of the corresponding observation belonging to the different clusters; each observation is assigned the cluster associated with the maximum probability in its cluster probability vector. We use φ_i to denote the cluster probability vector of the *i*-th observation. The learning task of variational inference is to find variational densities that maximize the ELBO; this can be accomplished through coordinate ascent variational inference (Bishop, 2006). Lim and Wang (2018) derive variational density updates for latent variables of the Dirichlet Process Gaussian Mixture (DPGM) as shown in lines 4, 5 and 8 of Algorithm 1. We use these updates for the variational densities of *μ, τ, c* and *v* in our DPHM model but we must update Bernoulli and Poisson parameters as well. We derived the coordinate-optimal updates for $q(p_{tk})$ and $q(\lambda_{tl})$ which represent the variational densities of the Bernoulli and Poisson parameters, respectively:

$$q(p_{tk}) \propto Beta\left(p_{tk} \middle| \gamma_k + \sum_i \varphi_{it}\, x_{ik}^{(b)},\ \delta_k + \sum_i \varphi_{it}\left(1 - x_{ik}^{(b)}\right)\right) \quad (3)$$

$$q(\lambda_{tl}) \propto Gamma\left(\lambda_{tl} \middle| \varepsilon_l + \sum_i \varphi_{it}\, x_{il}^{(p)},\ \zeta_l + \sum_i \varphi_{it}\right) \quad (4)$$

$x_{ik}^{(b)}$ is the *k*-th Bernoulli feature of the *i*-th observation and $x_{il}^{(p)}$ is the *l*-th Poisson feature of the *i*-th observation. φ_{it} is the probability of the *i*-th observation belonging to cluster *t*. We apply these updates for all $t = 1..T$, $k = 1..r_b$ and $l = 1..r_p$.

We derived a novel variational inference algorithm to fit DPHM models with must-link constraints, outlined in Algorithm 1. In the algorithm, $x_{ij}^{(g)}$ is the *j*-th Gaussian feature of the *i*-th observation and c_{-i} represents a vector of all cluster assignments except c_i. We define a must-link constraint as a set of indices for observations that must cluster together. While there are variational inference algorithms for semi-supervised classification models (Kingma et al., 2014), with finite numbers of classes, and Gibbs samplers for semi-supervised infinite mixture models (Vlachos et al., 2009), our work presents the first variational inference algorithm designed for semi-supervised infinite mixtures. In our algorithm, all observations in a particular must-link constraint are associated with a single cluster assignment variable, as shown in Figure 1, so they are all assigned the same cluster. In each iteration of coordinate ascent, our algorithm updates variational densities of global latent variables (*μ, τ, p, λ* and *v*) and uses their expectations to calculate cluster probability vectors. For each must-link constraint *m*, with

observations $X^m = \{x_i\}_{i \in m}$ and cluster assignment c^m (the single cluster assignment variable for all observations in m), we derived the coordinate-optimal variational density of c^m:

$$q(c^m) \propto \exp\left[\mathbb{E}_{-c^m}[\ln P(X^m, c^m | \mu, \tau, p, \lambda, c^{-m}, v)]\right] \quad (5)$$

where $\mathbb{E}_{-c^m}$ is the expectation with respect to all latent variables except c^m and c^{-m} is a vector of all cluster assignments except c^m. We obtain the cluster probability vector for the observations in m by evaluating the variational density over different clusters. Derivations for variational density updates are provided in the supplementary information.

Algorithm 1 Variational Inference for Semi-supervised DPHMs

Input: Data X, hyperparameters $\alpha, \nu, \rho, a, b, \gamma, \delta, \epsilon, \zeta$ and must-link constraints M
Output: Variational posterior $q(c, v, \mu, \tau, p, \lambda)$
Initialize: for $t = 1..T$
  $q(c_i = t) = \phi_{it}$ where ϕ_i is a random probability vector for $i = 1..n$
  $q(c^m = t) = \phi_{mt}$ where ϕ_m is a random probability vector for $m \in M$
  $q(v_t) \propto Beta(v_t | 1, \alpha)$
  $q(\mu_{tj}, \tau_{tj}) \propto NG(\mu_{tj}, \tau_{tj} | \nu_j, \rho_j, a_j, b_j)$ for $j = 1..r_g$
  $q(p_{tk}) \propto Beta(p_{tk} | \gamma_k, \delta_k)$ for $k = 1..r_b$
  $q(\lambda_{tl}) \propto Gamma(\lambda_{tl} | \epsilon_l, \zeta_l)$ for $l = 1..r_p$
while ELBO not converged
  Update variational densities for global latent variables. For $t = 1..T$
    $q(v_t) \propto Beta(v_t | 1 + \sum_{i=1}^n \phi_{it}, \alpha + \sum_{i=1}^n \sum_{s=t+1}^T \phi_{is})$
    for $j = 1..r_g$
      $q(\mu_{tj}, \tau_{tj}) \propto NG(\mu_{tj}, \tau_{tj} | \frac{\rho_j \nu_j + \sum_{i=1}^n \phi_{it} x_{ij}^{(g)}}{\rho_j + \sum_{i=1}^n \phi_{it}},$
        $\rho_j + \sum_{i=1}^n \phi_{it}, a_j + \frac{1}{2} \sum_{i=1}^n \phi_{it},$
        $b_j + \frac{1}{2}[\sum_{i=1}^n [\phi_{it}(x_{ij}^{(g)} - \mathbb{E}(\mu_{tj}))^2] + \rho_j(\mathbb{E}(\mu_{tj}) - \nu_j)^2])$
    for $k = 1..r_b$
      $q(p_{tk}) \propto Beta(p_{tk} | \gamma_k + \sum_{i=1}^n \phi_{it} x_{ik}^{(b)}, \delta_k + \sum_{i=1}^n \phi_{it}(1 - x_{ik}^{(b)}))$
    for $l = 1..r_p$
      $q(\lambda_{tl}) \propto Gamma(\lambda_{tl} | \epsilon_l + \sum_{i=1}^n \phi_{it} x_{il}^{(p)}, \zeta_l + \sum_{i=1}^n \phi_{it})$
  for any i not in a must-link constraint
    $q(c_i) \propto \exp(\mathbb{E}_{-c_i}[\ln P(x_i, c_i | \mu, \tau, p, \lambda, c_{-i}, v)])$
  for each $m \in M$
    $q(c^m) \propto \exp(\mathbb{E}_{-c^m}[\ln P(X^m, c^m | \mu, \tau, p, \lambda, c^{-m}, v)])$
  Calculate $ELBO[q(\mu, \tau, p, \lambda, c, v)]$
return $q(\mu, \tau, p, \lambda, c, v)$

Benchmarking

To show the advantage of the DPHM model in clustering heterogeneous data in unsupervised and semi-supervised settings, we benchmarked its performance against other clustering models on three synthetic datasets with 10, 25 and 50 clusters, respectively. Each dataset had 1000 observations with 10 Gaussian, 10 Bernoulli and 10 Poisson features. For each type of feature, there are 5 informative features that have cluster-specific distributions, and 5 nuisance features that have the same distribution for all clusters. Performance was measured with the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985).

We benchmarked the DPHM model against a finite Gaussian Mixture Model (GMM) with a diagonal covariance matrix, a DPGM model with a diagonal covariance matrix (using our 'dphmix' Python package), and k-means on the three synthetic datasets in an unsupervised setting. As the predictions of each model vary due to parameter initializations, we applied each model 10 times on each dataset with different initializations. The GMM and k-means were given the correct numbers of clusters for each dataset.

In the semi-supervised setting, we benchmarked the DPHM model against a semi-supervised GMM (Georgi et al., 2010) with a diagonal covariance matrix, a semi-supervised DPGM with a diagonal covariance matrix, and constrained k-means (Wagstaff et al., 2001). Each dataset was partitioned into 10 equal sized groups, and 10-fold cross-validation was applied to calculate ARIs for each model with fixed initializations. In each fold, one group was used as the training set and the remaining nine comprised the testing set. For a given fold, we create must-link constraints to force observations, in the training set of the fold, with the same ground-truth labels to cluster together; the semi-supervised models had to cluster observations in the testing set of the fold given the must-link constraints. Additional details about our benchmarking methods are provided in the supplementary information.

Sensitivity Analysis

We conducted sensitivity analysis to determine whether predictions, for the mouse heart dataset, are consistent across different sizes of training data. We trained a semi-supervised DPHM model to force all 109 VISTA enhancers (complete training set) to cluster together. Then, we trained semi-supervised DPHM models with must-link constraints on 10 to 100 VISTA enhancers (in intervals of 10). For each training set size, we randomly sampled 5 training sets, with replacement, from the 109 VISTA enhancers and applied a semi-supervised DPHM model for each sample. Finally, we applied an unsupervised DPHM model (without must-link constraints) on the mouse heart dataset. α was set to 1 for every model. For the semi-supervised models, we define predicted active enhancers (PAEs) as regions that cluster with the training set (excluding the VISTA enhancers in the training set). For the unsupervised model, we define PAEs as the regions in the cluster with the most VISTA enhancers (the VISTA enhancers are not included in the set of PAEs).

Motif Enrichment Analysis

We used MotEvo (Arnold et al., 2012) to compute posterior probabilities of TFs binding to the regions in the mouse heart dataset given the position weight matrices (PWMs) for their motifs, multi-sequence alignments (MSAs) and a phylogenetic model. We extracted the PWMs from the JASPAR 2016 database (Mathelier et al., 2016) and found that 559 redundant PWMs had positive posterior probabilities for at least one region in the mouse heart dataset. Details about the MSAs and phylogenetic model are provided in the supplementary information.

We applied lasso logistic regression to discover motifs that discriminate functional classes of enhancers from predicted inactive regions. Four regressions were trained to discriminate four classes, respectively, from a class of predicted inactive regions where posterior probabilities across the 559 motifs were used as features for each region. In a regression model, the coefficient of a motif quantifies its enrichment in the class of interest compared to the class of predicted inactive regions. The significance of each coefficient in each regression model was measured through a permutation test (1000 permutations). We identified enriched motifs for each class by taking the motifs with significant (permutation test $p < 0.05$) and positive coefficients for the class. For each class, we used GOrilla (Eden et al., 2009) to identify significantly ($q < 0.05$) enriched processes for TFs associated with their motifs. All TFs for the 559 motifs in the regressions were used as the background set for GOrilla.

Results

The DPHM model achieved higher ARIs than the GMM and DPGM on all synthetic datasets in unsupervised and semi-supervised settings, which shows the advantage of incorporating Bernoulli and Poisson distributions into the DPHM model. The performance of the DPHM model was comparable to k-means, which had the advantage of being initialized with the true number of clusters. The DPHM and k-means achieved perfect ARIs on the 10-cluster dataset (given the right initializations) and had similar ARIs on the other datasets; however, the DPHM model achieved higher maximum ARIs (across initializations) for the 25-cluster and 50-cluster datasets in the unsupervised and semi-supervised settings (Supplementary Figure 1).

We applied a semi-supervised DPHM model to the mouse heart dataset and trained the model on the complete training set. It discovered 47 PAEs [hereafter referred to as the semi-supervised DPHM's predicted active enhancers (SDPHM-PAEs)] and most displayed epigenetic marks (Ernst et al., 2011) of active enhancers (Figure 2A). Gene ontology (GO) enrichment analysis (section 2.1 of supplementary information) revealed that the nearest genes of the SDPHM-PAEs were significantly enriched for expression in heart ventricle during postnatal development in mouse ($q = 3.98 \times 10^{-2}$), suggesting that the semi-supervised model may predict heart-specific active enhancers. One SDPHM-PAE, located 1,322bp upstream from the transcription start site (TSS) of *Actc1* (a gene that transcribes cardiac alpha-actin), was previously functionally validated and confirmed to drive *Actc1* expression (Fleischmann et al., 1998). Sensitivity analysis showed that the DPHM model can predict over half of the SDPHM-PAEs with 70 or more VISTA enhancers in the must-link constraint (Figure 2B). Furthermore, the SDPHM-PAE near *Actc1* was predicted to be an active enhancer for at least 3/5 training set samples with 70 or more VISTA enhancers in the must-link constraint.

For comparison, we applied an unsupervised DPHM to the mouse heart dataset and found that the cluster with the most VISTA enhancers had 17 VISTA enhancers and 104 PAEs. These PAEs had some marks of active enhancers (Supplementary Figure 2) although their average H3K27 acetylation (K27ac) signal (minmax-scaled K27ac features averaged across time points and PAEs) was significantly (one-sided t-test $p = 8.81 \times 10^{-4}$) lower than the average K27ac signal of SDPHM-PAEs. The nearest genes of the PAEs from the unsupervised model were not

significantly (q < 0.05) enriched with any GO terms, suggesting that the unsupervised model could not predict enhancers near heart-specific genes. The results for the semi-supervised and unsupervised DPHM models indicate that the training data may allow the DPHM model to predict heart-specific enhancers with stronger activity.

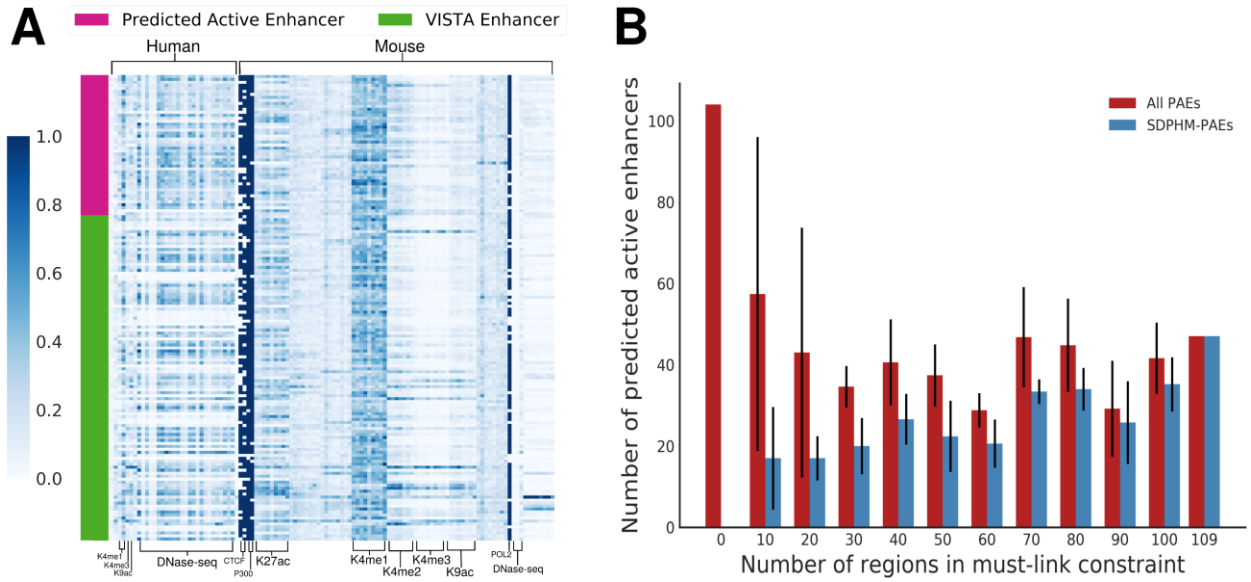

Fig. 2. (A) A heatmap comparing VISTA enhancers and SDPHM-PAEs from the semi-supervised DPHM model that was trained on the complete training set. Each column represents an ENCODE feature in heart tissues and features were minmax-scaled over the dataset to be between 0 and 1. (B) A comparison of the numbers of PAEs from the unsupervised DPHM model (with 0 regions in the must-link constraint) and semi-supervised DPHM models with differing training set sizes. For semi-supervised models with 10 to 100 regions in their training set, we randomly sampled their training sets, from the VISTA enhancers, 5 times and calculated the average number of PAEs, across the samples, for each training set size (red bars). In addition, for each training set size, we calculated the average number of PAEs that were also in the set of SDPHM-PAEs (blue bars). Standard deviations, for the numbers of PAEs and SDPHM-PAEs, across sampled training sets were also calculated (black lines).

Clustering the Clusters

The semi-supervised DPHM model that was trained on the complete training set found 47 clusters in the mouse heart dataset. In addition to the 47 SDPHM-PAEs, other clusters also contained regions with characteristics of active enhancers. Furthermore, groups of clusters shared characteristics of different enhancer states. We wanted to group similar clusters together using another clustering model. To this end, we represented each cluster (from the semi-supervised DPHM model that was trained on the complete training set) with a feature vector of expected values over its cluster-specific distribution parameters; all the values are continuous. We applied a DPGM ('dphmix' runs a DPGM when all features are continuous) with $\alpha = 1$ to group the clusters and found 14 classes (clusters of the original 47 clusters). There are five large (over 500 regions) classes that we number from 1 to 5 as shown in Figure 3A. We performed GO enrichment analysis on each class to identify possible functions of their regions.

Here, we analyze three interesting large classes (classes 2, 3 and 4). Analyses for the other two large classes are provided in the supplementary information.

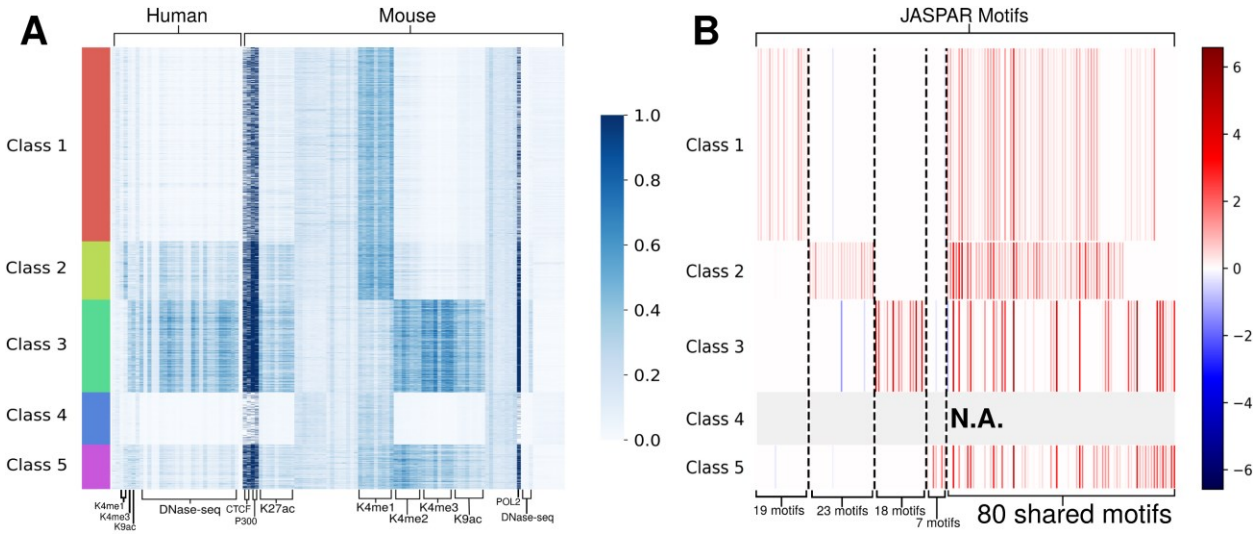

Fig. 3. (A) A heatmap showing the regions in the large classes. Each column represents an ENCODE feature in heart tissues and they were minmax-scaled to be between 0 and 1. (B) A heatmap showing the enrichment (regression coefficient) of JASPAR motifs in classes 1, 2, 3 and 5 compared to class 4 (predicted inactive regions). Only motifs that have a significant (permutation test $p<0.05$) and positive coefficient for at least one class compared to class 4 are included. Coefficients for class 4 are not available since it was used as the reference class for all the regressions.

We associated class 2 with active enhancers as it contains the cluster with all of the VISTA enhancers. It also contains 4 other clusters for a total of 741 regions. It has the highest average H3K4 mono-methylation (K4me1) signal (minmax-scaled mouse K4me1 features averaged across time points and regions in the class) and second highest average K27ac signal in embryonic mouse tissue among all classes. Consistent with typical active enhancers, P300 and POL2 were bound to the vast majority of the regions in adult tissue (Spicuglia and Vanhille, 2012). GO annotation analysis on genes near (within 1Mb) regions from the class revealed 159 significantly enriched biological processes including heart-specific processes like cardiac muscle hypertrophy ($q = 3.07 \times 10^{-5}$) and regulation of heart contraction ($q = 4.98 \times 10^{-5}$). To ensure that the observed enrichments were not biased by VISTA enhancers, we conducted GO enrichment analysis on the regions in the class that are not in VISTA. The regions were still significantly enriched with 13/159 of the aforementioned processes, suggesting that they may be novel active enhancers with similar functions as VISTA enhancers.

Class 3 is the second largest class and contains 1176 regions across 11 clusters. This class has the highest average H3K4 tri-methylation (K4me3), H3K9 acetylation (K9ac), DNase-seq signals (in both species) and K27ac in embryonic mouse tissue; all of these features are associated with active promoters (Ernst et al., 2011). Furthermore, over 96% of regions were bound to P300 and POL2 and 794 regions overlapped with TSSs of protein-coding genes. Another 224 were within 2kb of a TSS, suggesting that they could be promoters. GREAT indicated that the promoters of their nearest genes were significantly enriched with the Yy1

binding motif ($q = 6.22 \times 10^{-11}$); Yy1 (along with other TFs) binds to active enhancers and promoters to mediate interactions between them (Weintraub et al., 2017). The nearest (or overlapping) genes were enriched with 70 biological processes, none of which were specific to heart. Instead, they were enriched with DNA conformation change ($q = 2.26 \times 10^{-10}$) and general functions such as DNA metabolic process ($q = 5.16 \times 10^{-10}$), suggesting that Nkx2-5 binds to promoters of genes that are important for chromatin remodeling and house-keeping functions.

Class 4 only has one cluster of 663 regions and appears to largely contain inactive regions (Pradeepa, 2017) with very low K4me1 and DNase-seq signals (in both species) and K27ac in embryonic mouse tissue compared to the other enhancer classes (1 and 2). Furthermore, most regions were not bound to P300, CTCF or POL2 in adult tissue. The top two enriched biological processes were regulation of cardiac muscle cell differentiation ($q = 9.79 \times 10^{-4}$) and neuron differentiation ($q = 5.33 \times 10^{-3}$), suggesting that this class contains decommissioned enhancers that retained Nkx2-5 binding and played a role in heart and brain differentiation before deactivation.

Motif Enrichment Analysis of Classes

We performed motif enrichment analysis to identify differentially enriched motifs between classes. Each class was enriched with at least 60 different motifs, compared to class 4 (predicted inactive regions). In particular, 86 motifs were enriched in class 2 (the class that contains VISTA enhancers) and their TFs were significantly ($q < 0.05$) enriched for 17 biological processes, including regulation of muscle system process ($q = 6.59 \times 10^{-3}$). The motifs for other classes were not significantly enriched with any processes. While 80 motifs are enriched in multiple classes (shared motifs), each class is uniquely enriched with at least 7 motifs (Figure 3B), suggesting that many different TFs may control different enhancer states and multiple binding events may be required for enhancers to transition from an inactive state to a functional state.

Discussion

We derived the semi-supervised variational DPHM model to cluster heterogeneous data with must-link constraints using a Bayesian framework. We clustered genomic regions bound by a master regulator in embryonic mouse heart based on heterogeneous features and predicted the most similar regions to a set of known active enhancers. Furthermore, clustering the clusters revealed multiple functional classes of enhancers in the Nkx2-5 ChIP-seq dataset, and motif enrichment analysis suggested that many different motifs may be required to discriminate the classes from each other.

Our semi-supervised DPHM approach solves a different problem than other enhancer prediction methods (Ernst and Kellis, 2012; Hoffman et al., 2012; He et al., 2017; Li et al., 2018) as it predicts enhancer states from a set of bound regions. Our semi-supervised DPHM model

utilizes variational inference to learn from small training sets and handles continuous, binary and discrete features by incorporating heterogeneous exponential families. Even though the PAEs from the unsupervised DPHM model were weaker candidates for active enhancers compared to the SDPHM-PAEs, the unsupervised model discovered PAEs with some active enhancer marks so it can still be used for enhancer clustering. Gaussian, Bernoulli and Poisson distributions allowed the DPHM model to discover biologically relevant clusters, but features will not always follow these common distributions. Hence, the fit of the DPHM model may be improved by incorporating more distributions, such as exponentials or multinomials, which will require more benchmarking.

There are two computational advantages to using variational Bayesian methods for clustering tasks. First, the number of hyperparameters in an infinite mixture model only scales with the number of features, whereas, in a finite mixture model, it scales with the number of features and clusters. Our DPHM model has $1 + 4r_g + 2r_b + 2r_p$ hyperparameters, while a GMM, with a diagonal covariance matrix, would have $k - 1 + 2k(r_g + r_b + r_p)$, where k is the number of clusters. Hence, the DPHM model reduces the number of hyperparameters by a factor of order k, which can be substantial if k is large. Second, in the Bayesian framework, posterior probabilities can be estimated through Markov Chain Monte Carlo methods (Neal, 2000), however, convergence is difficult to assess and not guaranteed in a finite number of iterations. Variational inference provides a direct way to assess convergence through the ELBO function and is guaranteed to converge to a local optimum.

Further work is required to determine whether our DPHM model is viable for clustering regions genome-wide based on chromatin state. The algorithm is $O(\mathrm{Tn})$ as it passes through the entire dataset and all clusters to update cluster probability vectors, but would still require a large amount of time for large datasets as it has to complete a full pass over the dataset at each iteration of coordinate ascent. Moreover, there is substantial overhead in gathering files, extracting features for every genomic region and exhaustively testing different priors. Stochastic variational inference could speed up the algorithm for larger datasets as it allows some variational parameters to be updated based on subsamples of the data (Hoffman et al., 2013).

The purpose of clustering the clusters was to reduce the number of clusters, but, in principle, this can also be accomplished by tuning α. For large datasets, however, α has a marginal effect on the number of clusters, because variational densities for most of the stick-breaking variables are primarily determined by cluster probabilities rather than α (line 4 of Algorithm 1). The clustering of clusters is a novel, model-based method to hierarchically cluster data in a multi-way tree structure. Blundell *et al.* (2010) derived a hierarchical clustering algorithm that produces multi-way trees, but the algorithm is $O(\mathrm{n}^2 \log \mathrm{n})$, so it may have difficulties scaling to large data. In contrast, even for hierarchical clustering, our method is still $O(\mathrm{Tn})$, as each layer of clustering is done independently and only adds terms of order n to the time complexity. Nonetheless, our method is heuristic and further work is needed to create a principled, scalable Bayesian hierarchical clustering model.

Conclusion

In this paper, we derived the DPHM model to cluster genomic regions bound by a master regulator in embryonic mouse heart based on heterogeneous features. In addition, we derived a variational inference algorithm to force known active enhancers to cluster together. The semi-supervised DPHM model discovered 47 regions that were similar to the known active enhancers and near heart-specific genes. Furthermore, we clustered the clusters from the DPHM model to find 5 large classes of enhancers with distinct patterns over their features. The class with the known active enhancers was enriched with heart-specific biological processes and TF-binding motifs that are important for muscle system processes. The other functional classes of enhancers were also enriched with many different TF-binding motifs. Our results show that the semi-supervised DPHM model provides a principled Bayesian method for clustering heterogeneous data with small training sets.

Acknowledgements

We thank Rachel Chan, Lee Zamparo, Michael Hoffman and the members of the Moses Lab for helpful comments on the manuscript.

Funding

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant (AMM). JAM holds operating and infrastructure grants from NSERC, the Canada Foundation for Innovation, and the Ontario Ministry of Research and Innovation. Studentship funding to support this work was provided by Connaught International Scholarships (GS).

References

Arnold, P. et al. (2012). Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. Bioinformatics (Oxford, England), 28(4), 487–94.

Beal, M. (2003). Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer New York.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. Bayesian analysis, 1(1), 121–143.

Blundell, C. et al. (2010). Bayesian rose trees. In Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI'10, pages 65–72, Arlington, Virginia, United States. AUAI Press.

Bulger,M. and Groudine,M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell, 144(3), 327 – 339.

Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin: What, how, and why? Molecular Cell, 49(5), 825–837.
Catarino, R. R. and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes & Development, 32(3-4), 202–223.
Dupays, L. et al. (2015). Sequential binding of meis1 and nkx2-5 on the popdc2 gene: A mechanism for spatiotemporal regulation of enhancers during cardiogenesis. Cell Reports, 13(1), 183–195.
Eden, E. et al. (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. BMC Bioinformatics, 10(1), 48.
ENCODE Project Consortium (2012). An integrated encyclopedia of dna elements in the human genome. Nature, 489(7414), 57–74.
Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. Nature methods, 9(3), 215–6.
Ernst, J. et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), 43–49.
Fleischmann, M. et al. (1998). Cardiac specific expression of the green fluorescent protein during early murine embryonic development. FEBS letters, 440(3), 370–6.
Georgi, B. et al. (2010). Pymix - the python mixture package - a tool for clustering of heterogeneous biological data. BMC Bioinformatics, 11(1), 9.
He, Y. et al. (2017). Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proceedings of the National Academy of Sciences, 114(9), E1633–E1640.
Hinrichs, A. et al. (2006). The ucsc genome browser database: update 2006. Nucleic acids research, 34(suppl 1), D590–D598.
Hoffman,M.D. et al. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303–1347.
Hoffman, M. M. et al. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods, 9(5), 473–476.
Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218.
Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick breaking priors. Journal of the American Statistical Association, 96, 161–173.
Kingma, D. P. et al. (2014). Semi-supervised learning with deep generative model. In Advances in Neural Information Processing Systems, pages 3581–3589.
Li, Y. et al. (2018). Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics, 19(202).
Lim, K.-L. and Wang, H. (2018). Fast approximation of variational bayes dirichlet process mixture using the maximization–maximization algorithm. International Journal of Approximate Reasoning, 93, 153 – 177.
Mathelier, A. et al. (2016). Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 44(D1), D110–D115.
Neal, R.M. (2000). Markov chain sampling methods for dirichlet process mixture models. Journal of computational and graphical statistics, 9(2), 249–265.

Pradeepa, M. M. (2017). Causal role of histone acetylations in enhancer function. Transcription, 8(1), 40–47.
Spicuglia, S. and Vanhille, L. (2012). Chromatin signatures of active enhancers. Nucleus, 3(2), 126–131.
Visel, A. et al. (2007). Vista enhancer browser–a database of tissue-specific human enhancers. Nucleic acids research, 35(Database issue), D88–92.
Vlachos, A. et al. (2009). Unsupervised and constrained dirichlet process mixture models for verb clustering. In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, GEMS ’09, pages 74–82, Stroudsburg, PA, USA. Association for Computational Linguistics.
Wagstaff, K. et al. (2001). Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML ’01, pages 577–584, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
Weintraub, A. S. et al. (2017). Yy1 is a structural regulator of enhancer-promoter loops. Cell, 171(7), 1573–1588.e28.
Zentner, G. E. et al. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res., 21, 1273 – 1283.