Details of Chromatin Module INference on Trees (CMINT)

The CMINT framework aims to solve multiple clustering problems, one per cell type, simultaneously, by using a probabilistic generative model of the data. CMINT is similar to an existing approach we developed, Arboretum [1] for clustering multiple species, which we extend here to be applicable to cell lineages. Specifically to handle cell lineage data: (a) we handle arbitrary tree topologies unlike in Arboretum which can handle only a binary tree, (b) we extend the generative model to handle the observed data at the leaf and internal nodes of the tree. Below we describe the key aspects of the CMINT model and learning algorithm to estimate the parameters of this model.

CMINT generative model: The generative model must generate values for two types of random variables: (a) hidden variables representing the cluster assignments, and (b) observed variables encoding expression for each gene at all points in the tree. Let n denote the number of different chromatin mark datasets and τ denote the lineage tree describing how the cell types are related. Let k be the number of clusters for each cell type's chromatin mark dataset. The cluster assignments are matched between the datasets by the generative model of cluster assignments. CMINT's generative model is defined by the following components: $\{\mathbf{M}, \mathbf{S}, \pi, \mathbf{T}\}$, where $\mathbf{M} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_l, \cdots, \boldsymbol{\mu}_n\}$ and $\mathbf{S} = \{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_l, \cdots, \boldsymbol{\Sigma}_n\}$ correspond to the mean and covariances associated with the n cell type-specific chromatin mark datasets. π denotes a multinomial distribution for cluster assignments and $\mathbf{T}$ is the set of transition probability matrices for each branch on the tree defining the probability of genomic region to maintain or change its module assignment from its immediate predecessor cell state. Each $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ corresponds to the parameters defining the Gaussian mixture model for the l^{th} cell type, including k mean vectors $\boldsymbol{\mu}_l = \{\mu_{l1}, \cdots, \mu_{lk}\}$ and k covariances $\boldsymbol{\Sigma} = \{\Sigma_{l1}, \cdots, \Sigma_{lk}\}$ for the k Gaussian components of a GMM. Each transition probability matrix T_l for a cell type l that is not the root, is $k \times k$ where the entry $T_l(i, j)$ specifies the probability of a region to be in cluster i in node l of the tree, given that in its parent the gene is in cluster j. The generative model of CMINT generates the cluster assignments and chromatin profiles for each genomic region g in all n datasets in the following manner. We use l to denote the l^{th} node in the tree and $p(l)$ as the parent of node l. Let r denote the index of the root node.

- Set current node l to root, r.
- For each node l on tree τ
- if l is the root node, draw a cluster assignment for genomic region g by sampling from π, $k_g^l \sim \pi$
- else
  - j be the g's cluster assignment in l's parent
  - $k_g^c \sim T_l(j, :)$
  - Draw chromatin profile x_g^l for g in l using μ_{lk}, Σ_{lk}, where $k = k_g^l$.

At the end of this process, we have the chromatin profile for region g at each point in the tree τ. The CMINT algorithm uses the Expectation Maximization (EM) framework to infer all parameters. In the E-step, we infer the expected values of cluster assignments, $\gamma_{k|k'}^{li}$ of all genes at all the leaf nodes. We then recursively infer the expected values of cluster assignments of an intermediate node using the data nodes in the subtree below it and at this node. In the M-step, we estimate the mean and co-variance matrix of each cluster using $\gamma_{k|k'}^{li}$ for dataset c and also the transition probabilities. These steps are described below.

Expectation step: To infer the probability of a module assignment, we define k to be cluster assignment at one node and k' to be cluster assignment at the parent node. $\gamma^{li}_{k|k'}$ represents for posterior probability of the i^{th} genomic region at the l^{th} cell type to be in cluster k given that it is in cluster k' in the parent cell type of c. This will infer a $K \times K$ matrix $\Gamma^c(k,k') = \gamma^{li}_{k|k'}$. Then we infer in a recursive manner the posterior probabilities for all the intermediate nodes. In addition to this we also used α^{li} which is $K \times 1$ vector, with each element representing $\alpha^{li}(k) = \sum_{k'} \gamma^{li}_{k'|k}$ which says the probability of an observation given that the parent is in state k. The estimation is done recursively as follows for a region i at node l

- if l is a leaf
  1. $\gamma^{li}_{k|k'} = P_l(k|k')e_{i|\mu^l_k,\sigma^l_k}$
  2. $\alpha^{li}(k) = \sum_{k'} \gamma^{li}_{k|k'}$
  3. $\gamma^{li}_{k|k'} = \frac{\gamma^{li}_{k|k'}}{\alpha^{li}(k)}$

  where $P_l(k|k')$ is the conditional transition probability associated with cell type l. $e_{i|\mu^l_k,\sigma^l_k}$ corresponds to the probability of observing the i^{th} measurement from the k^{th} Gaussian component in cell type l.
- otherwise
  1. Estimate $\gamma^{l_c i}_{k''|k'}$ and $\alpha^{l_c i}$, for each child node l_c of l.
  2. $\gamma^{li}_{k|k'} = e_{i|\mu^l_k,\sigma^l_k} P_l(k|k') \prod_{l_c} \alpha^{l_c i}(k)$
  3. $\alpha^{li}(k) = \sum_{k'} \gamma^{li}_{k|k'}$

Maximization step: There are two sets of parameters in this model: (a) cluster transition probabilities, (b) Gaussian mixture model parameters. We assume that the co-variance matrix is a diagonal matrix. Each of these parameters can be estimated in closed form by deriving the expected likelihood with respect to the parameters. Specifically, the mean μ^l_k is estimated as

$$\mu^l_k = \frac{\sum_i \mathbf{x}_i \alpha^{li}(k)}{\sum_i \alpha^{li}(k)}$$

Variance for the m^{th} dimension is estimated as

$$\Sigma^l_k(m,m) = \frac{\sum_i (\mathbf{x}_i(m) - \mu^l_k(m))^2 \alpha^{li}(k)}{\sum_i \alpha^{li}(k)}$$

Transition probability for the l^{th} cell type is

$$T_l(k|k') = \frac{\sum_i \gamma^{li}_{k|k'}}{\sum_{k,k'} \gamma^{li}_{k|k'}}$$

CMINT model learning: We begin with an initial clustering assignment obtained from partitioning genes into k partitions, where k specifies the number of clusters. This partitioning can be obtained by randomly splitting the data, or by a clustering algorithm. We next repeat the expectation and maximization steps until convergence or until a fixed number of iterations have been executed.

References

[1] Sushmita Roy, Ilan Wapinski, Jenna Pfiffner, Courtney French, Amanda Socha, Jay Konieczka, Naomi Habib, Manolis Kellis, Dawn Thompson, and Aviv Regev. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Research*, 23(6):1039–1050, 2013.