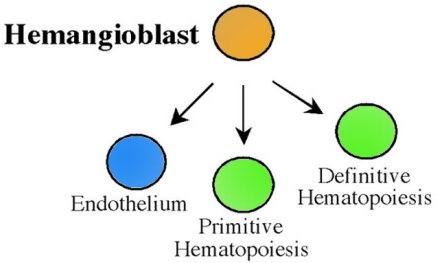
Hematopoietic Regulatory Network Inference

Jeremie Perrin, Corbin Hopper

ENS Cachan

November 18, 2019

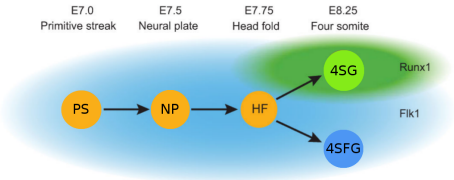
The Early Hematopoietic Process



Pluripotent cells differentiate into:

- Endothelial Cells
- Hematopoietic Cells

The Experiment

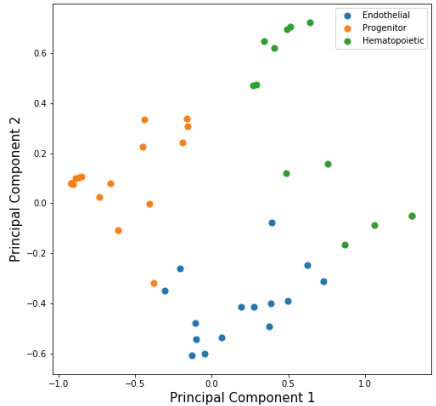


The data acquired:

- 3934 single cells
- 46 genes
- Binarized expression

Cell type	Number of embryos	Cells sorted	Cells retained	Percentage retained
PS	12	725	624	86.1
NP	9	637	552	86.7
HF	8	1,184	1,005	84.9
4SG	3	1,085	983	90.6
4SFG-	4	858	770	89.7
Total	36	4,489	3,934	87.6

Categorizing genes



Steps :

- Each gene is a 5D vector.
- Principal Component Analysis to reduce dimensions.
- K-means as unsupervised clustering.

Genes grouped as expected (known cell-type specific genes are grouped together).

Entropy Inference: Approach

- Maximum Likelihood Like
- Maximize mutual information $I(X;Y)$ of graph $G(V,E)$
- But many random distributions have some small probability of being correlated

$$\arg\max_{E \in G} \sum_{v \in V} I(v; e_{in}) - \mathbb{E}[I(A;B)] \tag{1}$$

Entropy Inference: Approach

- Optimization can consider in-edges to each node independently

$$\arg\max_{E \in G} \sum_{v \in V} I(v; e_{in}) - \mathbb{E}[I(A; B)] = \sum_{v \in V} \arg\max_{e_{in} \in E} I(v; e_{in}) - \mathbb{E}[I(A; B)]$$

- Optimize is independent of $p(v)$

$$= \sum_{v \in V} \arg\min_{e_{in} \in v} H(v|e_{in}) - \mathbb{E}[H(A|B)]$$

- Minimize the uncertainty of each gene over the set of genes that regulate it

Entropy Inference: Approach

How to simplify $\mathbb{E}[H(A|B)]$?

- Let a be a vector of length n, b an n by m matrix
- Assume a and columns of b are from the same set of random distributions
- As n approaches infinity, the probability that all possible 2^m vectors are in b approaches 1, and so:

$$\mathbb{E}[H(A|B)] \approx \log 2(m+1) - \log 2(m)$$

Entropy Inference: Algorithm

Infer Graph

Begin with a fully connected graph $G(V, E)$.
For each vertex $v \in V$:
    if $H(v) = 0$: remove v
For each vertex $v \in V$:
    Infer Node (v)

Infer Node (v)

Let w be the set of all predecessors of v, such that $(w_i, v) \in G$.
For each directed edge (u,v):
    Let w\u be the set w, excluding vertex u.
    If $H(v|w\backslash u) - H(v|w) \leq \mathbb{E}[H(Y|X) - H(Y|X\backslash x)]$:
        Remove edge (u,v)
        Infer Node (v)

Entropy Inference: Algorithm

$$\sum_{v \in V} \arg\min_{e_{in} \in v} H(v|e_{in}) - \mathbb{E}[H(A|B)] = \sum_{v \in V} \arg\min_{e_{in} \in v} H(v|e_{in}) - \log_2(\frac{m+1}{m})$$

Remove Edge if:

$$H(v|e_1, ..., e_{m-1}) - H(v|e_1, ..., e_m) \leq \mathbb{E}[H(a|b_1, ..., b_{m-1}) - H(a|b_1, ..., b_m)]$$

$$H(v|e_1, ..., e_{m-1}) - H(v|e_1, ..., e_m) \leq \log_2(\frac{m^2}{m^2-1}) \tag{2}$$

Entropy Inference: Algorithm

Why top down and not bottom up?

- Let there be a set of edges that collectively reduce $H(Y|X)$, but individually do not
- Example of XOR: $H(Y|x_1) = H(Y|x_2) = H(Y)$, but $H(Y|x_1, x_2) = 0$
- Top Down: removing any edge increases $H(Y|X)$, so none are removed
- Bottom up: algorithm checks edges one at a time, adds none, and stops

Entropy Inference: Performance

- Perfect on asymetric binary problems such as $Y = (x_1\ NAND\ x_2\ NAND\ x_3)\ AND\ (x_4\ OR\ x_5)$
- Undirected for symetric binary problems such as $Y = x_1\ XOR\ x_2$

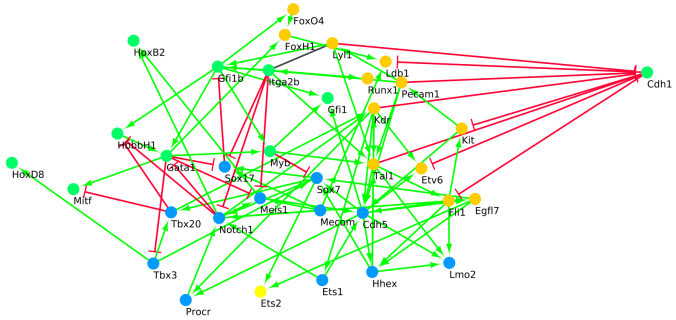
Entropy Inference: Performance

- Terrible on benchmark tests (BNLearn: Lizards, Coronary, Asia)
- Not robust to noise, poor performance on noisy binary problems
- Threshold to remove edges may be too lenient for $n < 2^m$, since $\mathbb{E}[H(A|B)]$ assumption would not hold

Entropy Inference: Hematopoietic Data

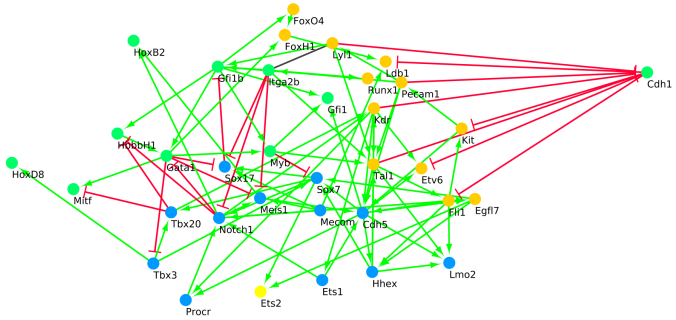
- 4 genes were removed since their entropy alone was 0 (housekeeping)
- Algorithm removes 7 more genes, reducing the total to 35
- Resulting average in-degree and out-degree is 28
- Since $n < 2^m$ entropy inference is too lenient and graph is too dense
- Instead used as preprocessing technique, followed by MIIC

Inferred Hematopoietic Regulatory Network



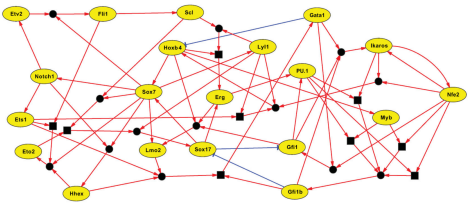
- Cdh1
  - High precursor expression levels
  - Clustered as hematopoietic
  - Biologically produces epithelial cadherin

Inferred Hematopoietic Regulatory Network



- Cdh5
  - Clustered as endothelial
  - Biologically linked to the process of endothelial development
  - Inferred hub with in-degree of 6 and an out-degree of 7

Their Network Inference Method



Their method :

1. Build state-transition graph
2. Reduce problem to SAT

Pros of their method :

- Explanatory
- General approach
  Given enough cells

Open question

How does one integrate an assumption of dynamicity into MIIC ?