Project: Bayesian Network for Stem Cell Differentiation **Lab: Beer**

Background: A Bayesian Network is a graphical way of representing probabilistic dependencies between variables in a system, and is a way to hypothesize causal relations between observable variables and the behavior of the system. In this project you will use a Bayesian Network to model how protein expression influences the differentiation of induced Pluripotent Stem Cells (iPSCs). iPSCs are similar to embryonic stem (ES) cells in that they are pluripotent (can differentiate into hundreds of different cell types), but in contrast to ES cells, pluripotent stem cells are artificially derived from a non-pluripotent adult cell, and therefore have possible immunological and ethical advantages compared to ES cells. In an important advance in regenerative medicine, iPSCs were first produced in 2006 from mouse cells and in 2007 from human cells in a series of experiments in Yamanaka's lab at Kyoto University.

The necessary signals for the creation of iPSCs and their differentiation into other cell types is not completely understood, but has been experimentally controlled by the expression of specific transcriptional regulatory proteins, including: OCT4, SOX2, NANOG, KLF4, MYC, REX1, PAX6, MEF2, and others. In this project you will find which combinations of these proteins determine the differentiation of iPSCs into neurons in a set of high throughput experiments. In this simple model, we will treat the expression level of each measured protein as a binary variable, i.e. OCT4=(0,1) and measured the cell differentiated state as a binary variable N=(0,1) where 0 is an iPSC and 1 is a differentiated neuron.

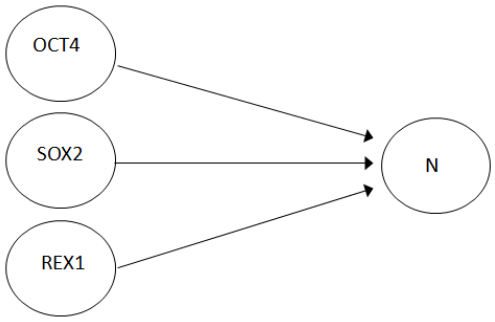
The file ipsc1.dat contains protein expression and cell state data for a large number of experiments. This simulated data was created by the rule:

> If (OCT4==1)&&(SOX2==1) then N=0 with P=0.9 and N=1 with P=0.1
> If (OCT4==0)&&(REX1==1) then N=0 with P=0.1 and N=1 with P=0.9
> and N=0 with P=0.5 otherwise.

In ipsc1.dat the first three columns represent OCT4, SOX2, and REX2, and the 7th column is N, but in general let's refer to the first 6 columns as X_1, X_2,...,X_6.

A Bayesian Network describes this rule graphically, as shown below on the right, where arrows represent conditional dependencies between variables. Here OCT4, SOX2, and REX1 affect the state of N. Equivalently, the rule can be described by the table:

if:	P(N=0)	P(N=1)
OCT4 and SOX2	0.9	0.1
!OCT4 and REX1	0.1	0.9
else	0.5	0.5



To write the conditional probability table more generally, since there are 3 variables affecting N, P(N) depends on the 2^3 states of the "parent nodes" OCT4, SOX2, and REX1 :

OCT4	SOX2	REX1	P(N=0)	P(N=1)
1	1	1	0.9	0.1
1	1	0	0.9	0.1
1	0	1	0.5	0.5
1	0	0	0.5	0.5
0	1	1	0.1	0.9
0	1	0	0.5	0.5
0	0	1	0.1	0.9
0	0	0	0.5	0.5

The likelihood that any Bayesian network describes the data follows from Bayes rule, and is given by:

$$P = P(n_p)\prod_{j=1}^{n_{states}} \frac{n_{0j}!n_{1j}!}{(n_{0j}+n_{1j})!} \quad \text{or} \quad \log_2 P = \sum_{j=1}^{n_{states}} \log_2\left(\frac{n_{0j}!n_{1j}!}{(n_{0j}+n_{1j})!}\right) - Cn_p ,$$

where j is an index over all parent nodes states, or rows in the conditional probability table. For the 3 parent node example above: j=0...7, where [0=(0,0,0), 1=(0,0,1), 2=(0,1,0), ... 7=(1,1,1)]. n_{0j} is the number of times N=0 in the data set when the parent node state is j, and n_{1j} is the number of times N=1 in the data set when the parent node state is j. The final term is a correction for network complexity: $P(n_p) = 2^{-C\,np}$, where n_p is the number of "parent" nodes in the network (here 3), and C is a constant depending on the size of the dataset, here use C=55. It is usually computationally more convenient to deal with log$P$ than P. For large n some terms need to be evalutated using Stirling's approximation:

$$\ln n! \approx \left(n+\frac{1}{2}\right)\ln n - n + \frac{1}{2}\ln(2\pi) \ .$$

Project:

1. Write a program to show that the maximum likelihood one parent node network for ipsc1.dat is OCT4→N with $\log_2 P$=-1373.52 by scoring all possible one parent node networks X_1→N, X_2→N,..., X_6→N and comparing their $\log_2$P. In this case the associated probability table is given by the observed frequencies: P(N=0)= $n_{0j}/(n_{0j}+n_{1j})$ and P(N=1)= $n_{1j}/(n_{0j}+n_{1j})$ for OCT4=1: (j=1) and OCT4=0: (j=0)

X_1	j	n_{0j}	n_{1j}	P(N=0)	P(N=1)
1	1	583	230	0.717	0.283
0	0	211	512	0.292	0.708

2. Write a program to score all possible networks with 1-6 parents and show that the maximum likelihood network now has $\log_2 P$=-1269.14. This network is therefore 2^{104} times more likely than the best one parent network. Show that the best inferred network corresponds to the rules used to generate the data:

X_1	X_2	X_3	j	n_{0j}	n_{1j}	P(N=0)	P(N=1)
1	1	1	7	184	28	0.868	0.132
1	1	0	6	192	16	0.923	0.077
1	0	1	5	98	96	0.505	0.495
1	0	0	4	109	90	0.548	0.452
0	1	1	3	14	149	0.086	0.914
0	1	0	2	88	99	0.471	0.529
0	0	1	1	20	155	0.114	0.886
0	0	0	0	89	109	0.449	0.551

This shows that Bayesian networks can identify the set of conditional dependencies in a dataset. In this example we found that OCT4 and SOX2 must be absent, and REX1 must be present, for an iPSC to differentiate into a neuron with high probability (if we didn't already know that).

3. Now read in ipsc2.dat, which was generated by different network. Find the highest scoring network and show your best estimate for the probability table that was used to generate the data.

Further reading:

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, and Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* (2007).
Cooper GF and Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* (1992).