**Introduction to Computational Bioengineering** **Mock Final Solutions**

**The Mock Final is open book, notes and articles. It is to be completed solo, without consulting others. Time limit is 45 min.**

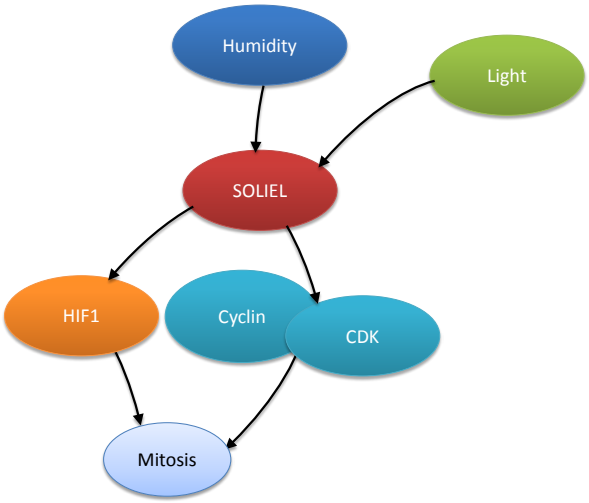
**Total Points: 100**

**Problem 1 (50 pts).** You're a first year graduate student collaborating with the Jet Propulsion Lab to analyze live cells that astrophysicists discovered on the surface of Mars in 2019. One day in the lab, you make an amazing discovery: when the humidity changes in San Antonio, most of the cells overexpress a light-sensitive compound that glows bright red. You name this compound SOLEIL. You observe that SOLIEL also changes how fast the cells grow. After a series of follow-up protein array studies, many known proteins from Earth eukaryotes show up, and you make the following conclusions:

SOLEIL is upstream of the cell cycle protein complex Cyclin B / Cyclin-Dependent Kinase 1 (Cyclin B-CDK1). This complex can coregulate cell mitosis. You also find that SOLIEL regulates a hypoxic response transcription factor (HIF1). HIF1 also regulates cell mitosis.

Excited by your discovery, you quickly sketch out a diagram of the relationships between the environmental stimuli , the three compounds (SOLIEL, Cyclin B-CDK1, HIF1), and mitosis.

a) Draw that diagram here, and state any assumptions:



b) Aiming for the Nobel Prize rather than the Ig Nobel, you decide you'll need to perform a series of rigorous tests to prove these relationships. To estimate the number of experiments and controls you'll need, you represent your diagram in (a) by Bayesian relationships.

(i) For simplicity, you also assume Boolean values for all your variables. If you assume conditional independence of variables in your diagram, how many probabilities/parameters do you need to calculate the joint probability of everything being on or present?

Conditional independence means the Bayesian signaling network assumption of being independent of grandparents (only dependent on a parent molecule/node in the network).

P(Humidity, Light, SOLIEL, HIF1, CYCLIN/CDK1 Complex, Mitosis) = P(Mitosis | CYCLIN/CDK1 Complex, HIF1)* P(CYCLIN/CDK1 Complex | SOLIEL )* P(HIF1 | SOLIEL)* P(SOLIEL | Light, Humidity)* P(Light)* P(Humidity) =
4 + 2 + 2 + 4 + 1 + 1 = 14

(ii) If you made no assumptions about conditional dependence or Bayesian assumptions, how many parameters would you need? Write out the full joint probability calculation.

P(Humidity, Light, SOLIEL, HIF1, CYCLIN/CDK1 Complex, Mitosis) = P(Mitosis | CYCLIN/CDK1 Complex, HIF1, SOLIEL, Light, Humidity)* P(CYCLIN/CDK1 Complex | HIF1, SOLIEL, Light, Humidity)* P(HIF1 | SOLIEL, Light, Humidity)* P(SOLIEL | Light, Humidity)* P( Light | Humidity)* P(Humidity)
= 32 + 16 + 8 + 4 + 2 + 1 = 63

c) You did it! Your paper is coming out in the famous new journal Alien Biotech in 2021. Reviewers were very enthusiastic and suggested new experiments. They want to see time course analysis. (10 pts)

(i) You decide to use modeling to help you plan the time course experiments. How would you define relationships in your system (diagram from part a) in a time-dependent manner? Write out the equations and assumptions.

The relationships in Figure 1 could be modeled with a set of ODEs with the assumptions of first order kinetics which would need to be tested experimentally

$d[SOLEIL]/dt = k1 \times Humidity - k2 \times [SOLEIL] - k3 \times [SOLEIL]$

$d[CycB/Cdk1]/dt = k2 \times [SOLEIL] + k4 \times [HIF1] - k5 \times [Cycb/Cdk1]$

$d[HIF1]/dt = k3 \times [SOLEIL] - k4 \times [HIF1] - k6 \times [HIF1]$

$dP_{Mitosis}/dt = k5 \times [CycB/Cdk1] + k6 \times [HIF1]$

**Problem 2 (50 pts).** Ten years later, you're head of the Mars National Research Center, which occupies the only major building on the planet. Recent news of a bacteria-born deadly epidemic on Earth is threatening to shut down any transport between Mars and Earth for the foreseeable future. You are asked to assess the risk and minimize the spread of the organism should a visitor from Earth bring the bacteria to your center. Your Center houses every human on the planet. It has a medical clinic and restaurant.

a) You gather your close senior team, and inform them that they are essential to preserving human life on Mars. You need to determine who is most at risk and find ways to minimize their potential exposure to the bacteria. They concur and quickly provide details on their potential high contact risk activities. List all assumptions in your analysis / calculations.

| Mars Senior Scientist | Works in Medical Clinic | Ave. Times / Day eats at Restaurant | Ave. Days Travels to Earth / Month | Ave. Days Contact with Other Scientists / Month |
|---|---|---|---|---|
| Andy | Yes | 1 | 4.20 | 28.00 |
| Bagrat | Yes | 2 | 7.58 | 17.00 |
| Caitlin | Yes | 3 | 5.37 | 14.00 |
| Calvin | No | 3 | 1.00 | 26.00 |
| Chelsey | No | 2 | 0.27 | 25.00 |
| Christopher | Yes | 3 | 5.00 | 17.00 |
| Emily T | No | 3 | 1.20 | 22.00 |
| Jacob | Yes | 2 | 0.00 | 30.00 |
| John | No | 4 | 0.74 | 25.00 |
| Maria | No | 4 | 2.01 | 12.00 |
| Nicholaus | Yes | 3 | 0.68 | 11.00 |
| Reid | No | 1 | 8.00 | 18.00 |
| Emily R | Yes | 2 | 2.12 | 29.00 |
| Ruojia | Yes | 1 | 8.40 | 26.00 |
| Samantha | No | 2 | 0.79 | 19.00 |
| Sydney | No | 4 | 3.87 | 20.00 |
| Tien | Yes | 1 | 8.55 | 16.00 |
| Yi | No | 3 | 7.00 | 24.00 |

Describe how you would categorize your team:

(i) Name two techniques that would allow you to quickly determine two major categories of scientists in the team: a higher and lower risk category.

k-means clustering
Hierarchical clustering with a set cut off

(ii) What metric would you use to compare team members?

Similarity by distance

(iii) How would you handle the categorical data?

Binarize the Yes / No into 1 and 0

(iv) Would you normalize the data? If so, how?

Yes, to the highest value in each column

(v) What is a technique that would allow you to determine which two scientists are the closest matches for risk of infection?

Hierarchical clustering

(vi) Which of the 4 listed activities are most likely to be grouped together? Why and how would you prove this?

Works in medical clinic, contact, times eats in restaurant – this is a hypothesis based on exposure to other people. It could be proven by clustering the activities using hierarchical or k-means, and see which grouped together.