BIOINF - MATH 540: Mathematics of Biological Networks

Assignment date: November 13, 2019

Due date: December 3, 2019

Name:

Problem Set 7: Evaluating Network Controllability from Data

Discrete-time-invariant linear control systems of the form

$$x[k+1] = Ax[k] + B_T u[k], \tag{1}$$

can be used to analyze control over the cell state, where $x[k]$ is the state at time k ($N \times 1$), A is the state transition matrix ($N \times N$), B_T is the input matrix ($N \times D_T$), and $u[k]$ is the input function ($N \times 1$).

In this assignment, the cell state refers to the type of cell (e.g. skin, muscle, etc) and we wish to reprogram the cell into a new type. We define each element in the state variable $x_i[k]$ to be the gene expression level of a Topologically Associating Domain (TAD) i at time k, which is defined as the sum of the expression levels of all genes contained within the TAD (units: fragments per kilobase of transcript per million, FPKM). TADs are inherent structural units of the chromosome: contiguous segments of the 1-D genome for which empirical physical interactions can be observed [1]. The initial cell type here will be fibroblast, and the set of initial snapshots $x^m_{\text{FIB}}, m = 1, ..., M$ represent the observed expression at the $M = 8$ different time points mentioned above [2]. A single arbitrary initial snapshot will be denoted x_I. A is derived from Dynamic Mode Decomposition (DMD) on the time series data (See: **Additional information**).

The input matrix B_T is defined from the regulatory set of T as follows: D_T = the number of TADs in the regulatory set of T, $\{i_1 < i_2 < ... < i_{D_T}\}$ = the indices of TADs in the regulatory set of T, column j of B_T have a 1-entry in row i_j, and zeros elsewhere.

The following measures are used to evaluate control:

1. **Minimal possible distance to target (μ_1)**
2. **Energy (μ_2)**
3. **Trace of controllability gramian ($\mathbf{tr}(W_c)$)**

Distance is defined as

$$d(u) = ||x[u, K, x_I] - x_F||_2 \tag{2}$$

with the standard Euclidean norm. The term $x[u, K, x_I]$ denotes the state of Equation 1 after K time steps, starting from state x_I and using input u. x_F is the target state. For each B_T and corresponding Equation 1, we can compute the optimal control $u^*(B_T)$ that minimizes the distance d to x_F. The energy of an input signal u is defined as

$$e(u) = \sum_{k=0}^{K-1} u[k]^T u[k]. \tag{3}$$

Both $d(u^*)$ and $e(u^*)$ quantify some capacity of Equation 1 to be steered from x_I to x_F. We define reprogramming measures 1 and 2 as

$$\mu_1(B) := d(u^*(B_T)) \tag{4}$$

$$\mu_2(B) := e(u^*(B_T)), \tag{5}$$

where smaller distances and energies imply higher capacity for reprogramming.

The controllability gramian W_c is defined as

$$W_c = \sum_{k=0}^{K-1} A^k B B^T (A^k)^T. \tag{6}$$

The trace of W_c is a measure of average controllability [3], where larger values imply greater control.

Problem 1

From the data provided, compute the following measures for each individual TF. For this problem, let $K = 3$, $x_I = x^8_{\text{FIB}}$, and $x_F = x_{\text{ESC}}$

(a) Minimal possible distance to target (μ_1)

(b) Energy of input that minimizes distance to target (μ_2)

(c) Trace of controllability gramian ($\text{tr}(W_c)$)

Problem 2

Where do the following TFs rank in each of the following measures: $\text{tr}(W_c)$ and μ_1 (out of the 332 TFs)?

(a) POU5F1 (b) SOX2 (c) KLF4 (d) MYC (e) NANOG

Answer for Problem 2
(a) POU5F1 μ_1 : 332; $\text{tr}(W_c)$: 1
(b) SOX2 μ_1 : 1; $\text{tr}(W_c)$: 2
(c) KLF4 μ_1 : 2; $\text{tr}(W_c)$: 3
(d) MYC μ_1 : 3; $\text{tr}(W_c)$: 4
(e) NANOG μ_1 : 4; $\text{tr}(W_c)$: 5

Additional information

MATLAB data variable explanation

- A: State transition matrix.
- X_fib: TAD-level RNA-seq expression time series data on fibroblasts (x_{FIB})
- X_esc: TAD-level RNA-seq expression for embryonic stem cells (x_{ESC})
- B: Control matrix
- TF_names: Names of transcription factors. These correspond to the columns of B

DMD computation

Recall from the main text we want to identify an A matrix such that $Y = AX$, where X and Y are the first and last $M-1$ columns of the initial data matrix x^m_{FIB}, respectively. A straight-forward identification of the matrix A is then computed by $A \approx \bar{A} := YX^\dagger$, where † represents the Moore-Penrose pseudoinverse. DMD computes the eigendecomposition $\bar{A}\Phi = \Phi\Lambda$ of the linear operator $\bar{A}$

Computation of control evaluations

The so-called controllability gramian W_c is an important characterizing matrix for any linear system 1, and is defined as

$$W_c = \sum_{k=0}^{K-1} A^k B B^T (A^k)^T, \tag{7}$$

The trace of W_c is a measure of the systems average controllability [3], where larger values imply greater

control. For comparison with our targeted control measures, we also compute $\mathrm{tr}(W_c)$:

$$\begin{aligned}\mathrm{tr}(W_c) &= \mathrm{tr}\left(\sum_{k=0}^{K-1} A^k B B^T (A^k)^T\right) \\ &= \mathrm{tr}\left(\sum_{k=0}^{K-1} B B^T (A^k)^T A^k\right) \\ &= \mathrm{tr}\left(B B^T \sum_{k=0}^{K-1} (A^k)^T A^k\right)\end{aligned}$$

That is, $\sum_{k=0}^{K-1} (A^k)^{\mathrm{T}} A^k$ can be computed once.

Computation of μ_1 and μ_4 is almost strictly linear algebra. For any (A, B), if A is $N \times N$ and B is $N \times D$, define the $N \times KD$ matrix C as

$$C = \left(B, AB, \ldots, A^{K-2}B, A^{K-1}B\right), \tag{8}$$

Finally, let

$$z := x_F - A^K x_I. \tag{9}$$

The minimal-distance control u_* is given by

$$u_* := C^\dagger z, \tag{10}$$

and $\mu_1(B)$ is the corresponding minimum distance given by

$$\mu_1(B) = \|Cu_* - z\| = \| \left(CC^\dagger - I_N\right) z\|_2. \tag{11}$$

$\mu_2(B)$ is the energy of u_*, i.e.

$$\mu_2(B) = e(u_*) = \sum_{k=0}^{K-1} u_*^{\mathrm{T}}[k] u_*[k] \tag{12}$$

References

[1] Jie Chen, Alfred O Hero III, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 2016.

[2] Scott Ronquist, Geoff Patterson, Lindsey A Muir, Stephen Lindsly, Haiming Chen, Markus Brown, Max S Wicha, Anthony Bloch, Roger Brockett, and Indika Rajapakse. Algorithm for cellular reprogramming. *Proceedings of the National Academy of Sciences*, 114(45):11832–11837, 2017.

[3] R.W. Brockett. *Finite Dimensional Linear Systems*. John Wiley & Sons, Inc., New York, USA, 1970.