BMI 203

Winter 2017

Homework Assignment #2

Due 02/07/2017 by 11:59pm PST

**In this exercise, you will be given a set of (~130) active sites to be clustered. You can retrieve the active sites and Python code that performs the I/O and contains some useful classes here:** [**https://github.com/ucsf-bmi-203-2017/hw2-skeleton**](https://github.com/ucsf-bmi-203-2017/hw2-skeleton)**.**

**Objective #1:** Implement a similarity metric to compare active sites.

* There is no “right” answer, but you should think hard about choosing something biologically meaningful.

**Objective #2:** Implement a partitioning algorithm to cluster the set of active sites.

**Objective #3:** Implement a hierarchical algorithm to cluster the set of active sites.

**Objective #4:** Implement a function to measure the quality of your clusterings.

**Objective #5:** Implement a function to compare your two clusterings

Please answer the following questions:

1. Explain your similarity metric, and why it makes sense biologically.
2. Explain your choice of partitioning algorithm.
3. Explain your choice of hierarchical algorithm.
4. Explain your choice of quality metric. How did your clusterings measure up?
5. Explain your function to compare clusterings. How similar were your two clusterings using this function?
6. Did your clusterings have any biological meaning?

**To complete this assignment:**

* Comment code: It is OK to get code from anywhere, but intuitive descriptions showing you understand what/why all steps are doing must be included.
* Email a single pdf (name = JaneSmith\_BMI203\_HW2.pdf) to [ryan.hernandez@ucsf.edu](mailto:ryan.hernandez@ucsf.edu) and [tamas@tamasnagy.com](mailto:tamas@tamasnagy.com) answering the above questions with prose and graphs.
* A link to your Github repository
  + - Make sure there is a link to the Travis build results for your repo in the README file
    - Note that only commits prior to the due date will be considered!