Python Lab

Pandas Library - II

Proteomics Informatics, Spring 2014

Week 8

18th Mar, 2014

Himanshu.Grover@nyumc.org

From last week…

Today

- Differential Gene Expression example
- leukemia dataset (Golub et. al.)
  - http://www.biolab.si/supp/bi-cancer/projections/

HW-3a

- For top-5 differentially expressed genes (according to p-value)
  - compute pairwise correlation matrix (it will actually be a data frame)
    - Check out the documentation of <DataFrame>.corr()
      - (exprDF.corr? in our example)
    - **Hint:**
      - Use exprDF.ix[:, <columns>] to get dataframe with only top-5 genes
      - Apply corr to this
  - Save it to another variable
  - Write it to a file:
    - Check out <DataFrame>.to_csv function

HW-3b

- For top-5 genes output summary statistics to a file
  - **Hint**:
    - Check out <DataFrame>.describe function
    - This outputs a dataframe. Save it to a variable and then to a file
    - Similar ideas from previous part

HW-3c

- Add a new function for preprocessing all gene expression values before applying t-test
- For every gene (x), perform:
  - (x-mean(x))/std(x, ddof=1)
  - Return this value
- Use apply() as discussed in the case of applying t-test
  - Note that this is similar to the “perform_tTest_two()” case, since this operation on each column returns back the whole column (Series) of values back.

References

- **Book:** Python for Data Analysis