DataCaptureAndScrub-Appendix

February 25, 2015

```
In [1]: import sys
        sys.version

Out[1]: '3.4.1 |Anaconda 2.1.0 (64-bit)| (default, Sep 10 2014, 17:10:18) \n[GCC 4.4.7 20120313 (Red Ha

In [2]: import numpy as np
        np.version.version

Out[2]: '1.9.0'

In [3]: import pandas as pd
        pd.version.version

Out[3]: '0.14.1'

In [4]: %timeit np.linalg.eigvals(np.random.rand(100,100))

100 loops, best of 3: 11.4 ms per loop
```

0.0.1 Automated data capture and cleaning

Assay data from:

Comprehansive Characterization of Cytochrome P450 Isozyme Selectivity across Chemical Libraries *Nat Biotechnology.* November ; 27(11): 1050-1055. doi:10.1038/nbt.1851

```
In [5]: # Download the molecular structures directly from PubChem
        !wget ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/.fetch/46/4074412786212022583.txt -P data/

--2014-06-02 18:17:58--  ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/.fetch/46/4074412786212022583.txt
           => 'data/4074412786212022583.txt'
Resolving ftp-private.ncbi.nlm.nih.gov (ftp-private.ncbi.nlm.nih.gov)... 130.14.29.30
Connecting to ftp-private.ncbi.nlm.nih.gov (ftp-private.ncbi.nlm.nih.gov)|130.14.29.30|:21... connected
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pubchem/.fetch/46 ... done.
==> SIZE 4074412786212022583.txt ... 1051404
==> PASV ... done.    ==> RETR 4074412786212022583.txt ... done.
Length: 1051404 (1.0M) (unauthoritative)

100%[======================================>] 1,051,404   --.-K/s   in 0.1s

2014-06-02 18:17:58 (8.93 MB/s) - 'data/4074412786212022583.txt' saved [1051404]

In [6]: # Rename file into something easier to work with.
        !cp data/4074412786212022583.txt data/1851smiles.txt
```

```
In [7]: # Load the structure data into a DataFrame
        smiles = pd.read_table('data/1851smiles.txt', sep='\t', names=('SID', 'SMILES'))

In [8]: # Inspect the first five rows of the DataFrame.
        smiles.head(5)
```

Out[8]:

	SID	SMILES
0	26751441	COC(=O)NC/C=C\C1=NC(=CO1)CCCO
1	26751440	C[C@H](C1=CC=CC=C1)N2C(=O)[C@@H]3CC[C@H]4[C@H]...
2	26751439	CCN1C(=O)[C@H]2CC=C3[C@H]([C@H]2C1=O)[C@@H]([C...
3	26751438	C1COC2([C@@H]3[C@H](O3)[C@H]([C@@H]4C2=CC[C@H]...
4	26751437	C=CC1=C[C@H]([C@H]2[C@@H](C13OCCCO3)O2)O

[5 rows x 2 columns]

```
In [9]: # Download bioassy data directly from PubChem
        !wget ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/.fetch/3/3837744530006510797.csv -P data/

--2014-06-02 18:18:11--  ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/.fetch/3/3837744530006510797.csv
           => 'data/3837744530006510797.csv'
Resolving ftp-private.ncbi.nlm.nih.gov (ftp-private.ncbi.nlm.nih.gov)... 130.14.29.30
Connecting to ftp-private.ncbi.nlm.nih.gov (ftp-private.ncbi.nlm.nih.gov)|130.14.29.30|:21... connected
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pubchem/.fetch/3 ... done.
==> SIZE 3837744530006510797.csv ... 12687619
==> PASV ... done.    ==> RETR 3837744530006510797.csv ... done.
Length: 12687619 (12M) (unauthoritative)

100%[=====================================>] 12,687,619  21.3MB/s   in 0.6s

2014-06-02 18:18:12 (21.3 MB/s) - 'data/3837744530006510797.csv' saved [12687619]

In [10]: # Rename file into something easier to work with.
         !cp data/3837744530006510797.csv data/1851bioassay.csv

In [11]: # Load the data into a DataFrame
         bioassay = pd.read_csv('data/1851bioassay.csv')

/home/ubuntu/anaconda3/lib/python3.4/site-packages/pandas/io/parsers.py:1070: DtypeWarning: Columns (12
  data = self._reader.read(nrows)

In [12]: # Inspect the first three rows of the DataFrame.
         bioassay.head(3)
```

Out[12]:

	#	SID	CID	BioAssay_Source	RankScore	Outcome	Xref	URL	\
0	1	26751441	10847630	NCGC	NaN	Unspecified	NaN	NaN	
1	2	26751440	16758818	NCGC	NaN	Unspecified	NaN	NaN	
2	3	26751439	16758817	NCGC	NaN	Unspecified	NaN	NaN	

	Comment	DepositDate	Inhibition	Observed	Approved Drug Collection	\
0	NaN	2009/07/08	False	Exploratory	NaN	
1	NaN	2009/07/08	False	Exploratory	NaN	
2	NaN	2009/07/08	True	Exploratory	NaN	

	Analysis Comment	Activity Outcome	Activity Score	\

```
0               NaN                 1               0
1               NaN                 1               0
2               NaN                 2              44

   p450-cyp2c19-Potency_&#956;M   p450-cyp2c19-Curve_Description  \
0                           NaN                             NaN
1                           NaN                             NaN
2                       1.58489  Partial curve; partial efficacy

   p450-cyp2c19-Fit_LogAC50  p450-cyp2c19-Fit_HillSlope
0                       NaN                        NaN ...
1                       NaN                        NaN ...
2                      -5.8                      3.132 ...

[3 rows x 150 columns]
```

In [13]:
```
# Inspect the size of the data set
bioassay.shape
```

Out[13]: (17143, 150)

0.1 Join Dataframes

In [14]:
```
merged = pd.merge(smiles, bioassay, on='SID')
merged.head(3)
```

Out[14]:
```
        SID                                             SMILES  #       CID  \
0  26751441                        COC(=O)NC/C=C\C1=NC(=CO1)CCCO  1  10847630
1  26751440  C[C@H](C1=CC=CC=C1)N2C(=O)[C@@H]3CC[C@H]4[C@H]...  2  16758818
2  26751439  CCN1C(=O)[C@H]2CC=C3[C@H]([C@H]2C1=O)[C@@H]([C...  3  16758817

  BioAssay_Source  RankScore      Outcome  Xref  URL  Comment DepositDate  \
0            NCGC        NaN  Unspecified   NaN  NaN      NaN  2009/07/08
1            NCGC        NaN  Unspecified   NaN  NaN      NaN  2009/07/08
2            NCGC        NaN  Unspecified   NaN  NaN      NaN  2009/07/08

  Inhibition Observed Approved Drug Collection  Analysis Comment  \
0              False   Exploratory        NaN               NaN
1              False   Exploratory        NaN               NaN
2               True   Exploratory        NaN               NaN

   Activity Outcome  Activity Score  p450-cyp2c19-Potency_&#956;M  \
0                 1               0                           NaN
1                 1               0                           NaN
2                 2              44                       1.58489

    p450-cyp2c19-Curve_Description  p450-cyp2c19-Fit_LogAC50
0                              NaN                       NaN ...
1                              NaN                       NaN ...
2  Partial curve; partial efficacy                      -5.8 ...

[3 rows x 151 columns]
```

0.2 Generate Descriptors in MOE and upload file.

In [16]:
```
descriptors = pd.read_csv('data/SIDWashedStructureDescriptors.csv')
```

```
In [17]: descriptors.head(3)
```

```
Out[17]:       SID                                        WashedMols       apol  \
         0  842238              Clc1cc(NC(=O)CCC)ccc1N1CC[NH+](CC1)CC  51.111824
         1  842250  Fc1ccc(cc1)Cn1nnnc1C[NH+](CC1=Cc2c(NC1=O)cc1OC...  66.848030
         2  842319  S1\C(=C/2\c3c(N(CC)C\2=O)cccc3)\C(=O)N(CCCOC)C1=S  52.328274

            a_acc  a_acid  a_aro  a_base  a_count  a_don  a_heavy  a_hyd       a_IC  \
         0      1       0      6       1       46      1       21     16   69.232559
         1      7       0     17       1       58      2       34     21  102.172100
         2      4       0      6       0       42      0       24     15   73.176926

               a_ICM  a_nB  a_nBr  a_nC  a_nCl  a_nF  a_nH  a_nI
         0  1.505056     0      0    16      1     0    25     0 ...
         1  1.761588     0      0    23      0     1    24     0 ...
         2  1.742308     0      0    17      0     0    18     0 ...

         [3 rows x 188 columns]
```

```
In []:
```

```
In [18]: mergedmore = pd.merge(merged, descriptors, on='SID')
```

```
In [20]: mergedmore.shape
```

```
Out[20]: (17143, 338)
```

```
In [21]: mergedmore.to_csv("data/complete.csv")
```

```
In [22]: columns = mergedmore.columns
```

0.3 Subset 2c19 data

```
In [24]: isozyme2c19 = mergedmore.reindex(columns=['SID','Activity Score','apol','zagreb'])
         isozyme2c19.head(10)
```

```
Out[24]:         SID  Activity Score       apol  zagreb
         0  26751441               0  35.436687      74
         1  26751440               0  55.646240     146
         2  26751439              44  49.834652     146
         3  26751438               0  59.968239     174
         4  26751437              10  31.903103      86
         5  26751436               0  40.283482     106
         6  26751435               0  51.324238     130
         7  26751434               0  46.470654     120
         8  26751433               0  49.834652     128
         9  26751432              10  53.510654     140

         [10 rows x 4 columns]
```

```
In [25]: isoz2c19 = mergedmore.loc[:,'apol':'zagreb']
         isoz2c19.head(5)
```

```
Out[25]:         apol  a_acc  a_acid  a_aro  a_base  a_count  a_don  a_heavy  a_hyd  \
         0  35.436687      3       0      5       0       33      2       17      8
         1  55.646240      5       0      6       0       49      2       26     17
```

```
2  49.834652      6       0      0       0       45      1       24     14
3  59.968239      6       0      6       0       52      1       29     19
4  31.903103      4       0      0       0       29      1       15     10

        a_IC     a_ICM  a_nB  a_nBr  a_nC  a_nCl  a_nF  a_nH  a_nI  a_nN  \
0  54.411259  1.648826     0      0    11      0     0    16     0     2
1  73.030655  1.490422     0      0    20      0     0    23     0     1
2  69.898079  1.553291     0      0    17      0     0    21     0     1
3  78.763672  1.514686     0      0    22      0     0    23     0     1
4  41.524731  1.431887     0      0    11      0     0    14     0     0

   a_nO
0     4 ...
1     5 ...
2     6 ...
3     6 ...
4     4 ...

[5 rows x 186 columns]
```

```
In [26]: isozy2c19 = mergedmore.ix[:,['SID','Activity Score', 1]]
         isozy2c19.head(3)
```

```
Out[26]:        SID  Activity Score   1
         0  26751441               0 NaN
         1  26751440               0 NaN
         2  26751439              44 NaN

         [3 rows x 3 columns]
```

```
In [27]: isozym2c19 = mergedmore.ix[:,'apol':'zagreb']
         isozym2c19.head(3)
```

```
Out[27]:        apol  a_acc  a_acid  a_aro  a_base  a_count  a_don  a_heavy  a_hyd  \
         0  35.436687      3       0      5       0       33      2       17      8
         1  55.646240      5       0      6       0       49      2       26     17
         2  49.834652      6       0      0       0       45      1       24     14

                 a_IC     a_ICM  a_nB  a_nBr  a_nC  a_nCl  a_nF  a_nH  a_nI  a_nN  \
         0  54.411259  1.648826     0      0    11      0     0    16     0     2
         1  73.030655  1.490422     0      0    20      0     0    23     0     1
         2  69.898079  1.553291     0      0    17      0     0    21     0     1

            a_nO
         0     4 ...
         1     5 ...
         2     6 ...

         [3 rows x 186 columns]
```

```
In [31]: # Load the data into a DataFrame
         isozyme2c19 = pd.read_csv('data/2c19.csv')
```

```
In [32]: isozyme2c19.head(5)
```

Out[32]:

	SID	p450-cyp2c19-ActivityScore	apol	a_acc	a_acid	a_aro
0	842238	0	51.111824	1	0	6
1	842250	20	66.848030	7	0	17
2	842319	20	52.328274	4	0	6
3	842408	90	42.691135	4	0	11
4	842584	41	36.787930	3	0	17

	a_base	a_count	a_don	a_heavy	a_hyd	a_IC	a_ICM	a_nB	a_nBr
0	1	46	1	21	16	69.232559	1.505056	0	0
1	1	58	2	34	21	102.172100	1.761588	0	0
2	0	42	0	24	15	73.176926	1.742308	0	0
3	0	31	0	22	13	60.580517	1.954210	0	0
4	0	28	2	18	13	43.328800	1.547457	0	0

	a_nC	a_nCl	a_nF	a_nH	a_nI	
0	16	1	0	25	0	...
1	23	0	1	24	0	...
2	17	0	0	18	0	...
3	13	0	0	9	0	...
4	14	1	0	10	0	...

[5 rows x 188 columns]