FINAL PROJECT

November 16, 2019

Classifying a protein as either acidic or basic using Instability index scores and Structure molecular weight in Daltons

In [70]:
```
#Libraries
import pandas as pd
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
import matplotlib.pyplot as plt
```

In [71]:
```
#Retrieve Data
df = pd.read_csv("pdb_test_plot_data.csv")
```

In [72]:
```
#Look at Data
df.head()
```

Out[72]:

	Unnamed: 0	structureId	classification	experimentalTechnique	macromoleculeType	residueCount	resolution	structureMolecularWeight	crystallizationMethod
0	114	1914	ALU DOMAIN	X-RAY DIFFRACTION	Protein	232	2.53	26562.88	hanging drop
1	129	1A04	SIGNAL TRANSDUCTION PROTEIN	X-RAY DIFFRACTION	Protein	430	2.20	47657.57	vapor diffusion - sitting drop
2	191	1A2B	ONCOGENE PROTEIN	X-RAY DIFFRACTION	Protein	182	2.40	21160.31	vapor diffusion - hanging drop
3	210	1A2X	COMPLEX (SKELETAL MUSCLE/MUSCLE PROTEIN)	X-RAY DIFFRACTION	Protein	206	2.30	23608.31	vapor diffusion - hanging drop
4	277	1A52	RECEPTOR	X-RAY DIFFRACTION	Protein	516	2.80	60742.53	vapor diffusion - hanging drop

```
   crystallizationTempK    ...     publicationYear  \
0                 277.0    ...              1997.0
1                 277.0    ...              1998.0
2                 277.0    ...              1998.0
3                 289.0    ...              1998.0
4                 291.0    ...              1998.0

                                              sequence  \
0  MASMTGGQQMGRIPGNSPRMVLLESEQFLTELTRLFQKCRSSGSVF...
1  SNQEPATILLIDDHPMLRTGVKQLISMAPDITVVGEASNGEQGIEL...
2  SMAAIRKKLVIVGDVACGKTCLLIVFSKDQFPEVYVPTVFENYVAD...
3  TDQQAEARSYLSEEMIAEFKAAFDMFDADGGGDISVKELGTVMRML...
4  MIKRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASM...

                                    Protein Analysis  Aromaticity  \
0  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.056034
1  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.018605
2  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.076923
3  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.063107
4  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.058140

   Instability Index        pI     Gravy   Helix %    Turn %   Sheet %
0          40.792241  9.637390 -0.672414  0.245690  0.215517  0.262931
1          33.597674  5.729553 -0.162791  0.297674  0.186047  0.358140
2          51.445055  4.986389 -0.325275  0.296703  0.181319  0.269231
3          51.867476  4.358093 -0.765534  0.223301  0.140777  0.378641
4          41.076744  6.510803  0.080620  0.344961  0.182171  0.368217

[5 rows x 24 columns]
```

In [73]:
```
#Data cleaning
proteins = df.copy()
proteins.drop(['classification', 'experimentalTechnique', 'macromoleculeType', 'residu
proteins.tail()
```

Out[73]:
```
       Unnamed: 0 structureId classification experimentalTechnique  \
54269      118238        5WZ1    TRANSFERASE     X-RAY DIFFRACTION
54270      118239        5WZ2    TRANSFERASE     X-RAY DIFFRACTION
54271      118240        5WZ3    TRANSFERASE     X-RAY DIFFRACTION
54272      118243        5X3A      HYDROLASE     X-RAY DIFFRACTION
54273      118247        5X6N  CELL INVASION     X-RAY DIFFRACTION

      macromoleculeType  residueCount  resolution  structureMolecularWeight  \
54269           Protein          2208        2.51                 248715.53
54270           Protein           828        2.60                  95630.65
54271           Protein           619        1.80                  72205.86
54272           Protein           756        1.79                  83489.55
```

```
54273          Protein           338        3.00                  40390.20

              crystallizationMethod  crystallizationTempK    ...     \
54269  VAPOR DIFFUSION, SITTING DROP                 291.0    ...
54270  VAPOR DIFFUSION, SITTING DROP                 291.0    ...
54271  VAPOR DIFFUSION, SITTING DROP                 291.0    ...
54272  VAPOR DIFFUSION, HANGING DROP                 287.0    ...
54273  VAPOR DIFFUSION, HANGING DROP                 293.0    ...

       publicationYear                                          sequence  \
54269           2017.0  HHHHHHGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRAL...
54270           2017.0  HHHHHHGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRAL...
54271           2017.0  HHHHHHMKIIGNRIERIRSEHAETWFFDENHPYRTWAYHGSYEAPT...
54272           2017.0  APNKPFPQHTTYTSGSIKPNHVTQSAMDNSVKAKWDSWKSAYLKTA...
54273           2006.0  MVINQTFLQNNVMDKCNDKRKRGERDWDCPAEKDICISDRRYQLCM...

                                      Protein Analysis  Aromaticity  \
54269  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.076087
54270  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.076087
54271  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.096931
54272  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.129630
54273  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...     0.094675

       Instability Index        pI     Gravy   Helix %    Turn %   Sheet %
54269          46.633913  9.176086 -0.458696  0.275362  0.253623  0.253623
54270          46.626365  9.157776 -0.458696  0.275362  0.253623  0.253623
54271          37.768336  8.409363 -0.612763  0.289176  0.198708  0.247173
54272          26.376481  5.962219 -0.526984  0.272487  0.320106  0.193122
54273          43.545266  8.960876 -0.809172  0.278107  0.198225  0.227811

[5 rows x 24 columns]
```

In [74]:
```
#If phValue is <7, the protein is acidic. If ph>=7, the protein is basic
#acids marked 0, bases marked 1
proteins['phCat'] = (proteins['pI'] >= 7)*1
proteins.head()
```

Out[74]:
```
   Unnamed: 0 structureId                             classification  \
0         114        1914                                ALU DOMAIN
1         129        1A04                SIGNAL TRANSDUCTION PROTEIN
2         191        1A2B                            ONCOGENE PROTEIN
3         210        1A2X  COMPLEX (SKELETAL MUSCLE/MUSCLE PROTEIN)
4         277        1A52                                 RECEPTOR

  experimentalTechnique macromoleculeType  residueCount  resolution  \
0     X-RAY DIFFRACTION           Protein           232        2.53
1     X-RAY DIFFRACTION           Protein           430        2.20
2     X-RAY DIFFRACTION           Protein           182        2.40
```

```
3     X-RAY DIFFRACTION           Protein           206        2.30
4     X-RAY DIFFRACTION           Protein           516        2.80

   structureMolecularWeight           crystallizationMethod  \
0                  26562.88                     hanging drop
1                  47657.57  vapor diffusion - sitting drop
2                  21160.31  vapor diffusion - hanging drop
3                  23608.31  vapor diffusion - hanging drop
4                  60742.53  vapor diffusion - hanging drop

   crystallizationTempK  ...    \
0                 277.0  ...
1                 277.0  ...
2                 277.0  ...
3                 289.0  ...
4                 291.0  ...

                                              sequence  \
0  MASMTGGQQMGRIPGNSPRMVLLESEQFLTELTRLFQKCRSSGSVF...
1  SNQEPATILLIDDHPMLRTGVKQLISMAPDITVVGEASNGEQGIEL...
2  SMAAIRKKLVIVGDVACGKTCLLIVFSKDQFPEVYVPTVFENYVAD...
3  TDQQAEARSYLSEEMIAEFKAAFDMFDADGGGDISVKELGTVMRML...
4  MIKRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASM...

                                 Protein Analysis Aromaticity  \
0  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...    0.056034
1  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...    0.018605
2  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...    0.076923
3  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...    0.063107
4  <Bio.SeqUtils.ProtParam.ProteinAnalysis object...    0.058140

   Instability Index        pI     Gravy   Helix %    Turn %   Sheet %  phCat
0          40.792241  9.637390 -0.672414  0.245690  0.215517  0.262931      1
1          33.597674  5.729553 -0.162791  0.297674  0.186047  0.358140      0
2          51.445055  4.986389 -0.325275  0.296703  0.181319  0.269231      0
3          51.867476  4.358093 -0.765534  0.223301  0.140777  0.378641      0
4          41.076744  6.510803  0.080620  0.344961  0.182171  0.368217      0

[5 rows x 25 columns]
```

In [75]:
```
#set y as ph category, set x as molecular weight and instability index
y=proteins[['phCat']].copy()
y.head()
proteinFeatures= ['structureMolecularWeight','Instability Index']
X=proteins[proteinFeatures].copy()
X.head()
```

Out[75]:
```
   structureMolecularWeight  Instability Index
0                  26562.88          40.792241
```

```
1                   47657.57          33.597674
2                   21160.31          51.445055
3                   23608.31          51.867476
4                   60742.53          41.076744
```

In [76]:
```
#Begin Testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state
ph_classifier = DecisionTreeClassifier(random_state=0)
ph_classifier.fit(X_train, y_train)
type(ph_classifier)
```

Out[76]:
```
sklearn.tree.tree.DecisionTreeClassifier
```

In [79]:
```
predictions = ph_classifier.predict(X_test)
```

In [80]:
```
predictions[:10]
```

Out[80]:
```
array([1, 0, 0, 0, 0, 0, 1, 0, 0, 0], dtype=int32)
```

In [81]:
```
y_test['phCat'][:10]
```

Out[81]:
```
5373     0
11724    0
11884    0
5491     0
41349    1
24926    0
27126    1
30492    0
34412    0
48303    0
Name: phCat, dtype: int32
```

In [82]:
```
accuracy_score(y_true = y_test, y_pred = predictions)
```

Out[82]:
```
0.74540784992462727
```

Predict Protein pH with Python

Junah Park

No need to request edit permission. To make a copy for your presentation, use “File->Download”.

Abstract

I am using the protein database. Could we create a classification model to predict whether a protein is an acid or a base given the protein's structure molecular weight in Daltons and its instability index? I used sklearn to create a classification model that would predict whether a protein is acidic or basic with roughly 75% accuracy

Motivation

Proteins are our building blocks; as in, the building blocks used to build us. A substance’s ph level greatly affects how it behaves. Having any model that can reliably identify that quality given other attributes would be helpful in a wide range of medicinal and research applications.

Dataset(s)

I am using the protein database; focusing on the structural molecular weight, instability index, and pI (isoelectric point) features.

Data Preparation and Cleaning

I needed to isolate which features I needed to use. I used the pI to separate proteins into 2 categories of acids and bases marked as 0 and 1. I thought that I would need a third category for neutral, but found that the protein database had included pI=7 in the bases category.

Research Question(s)

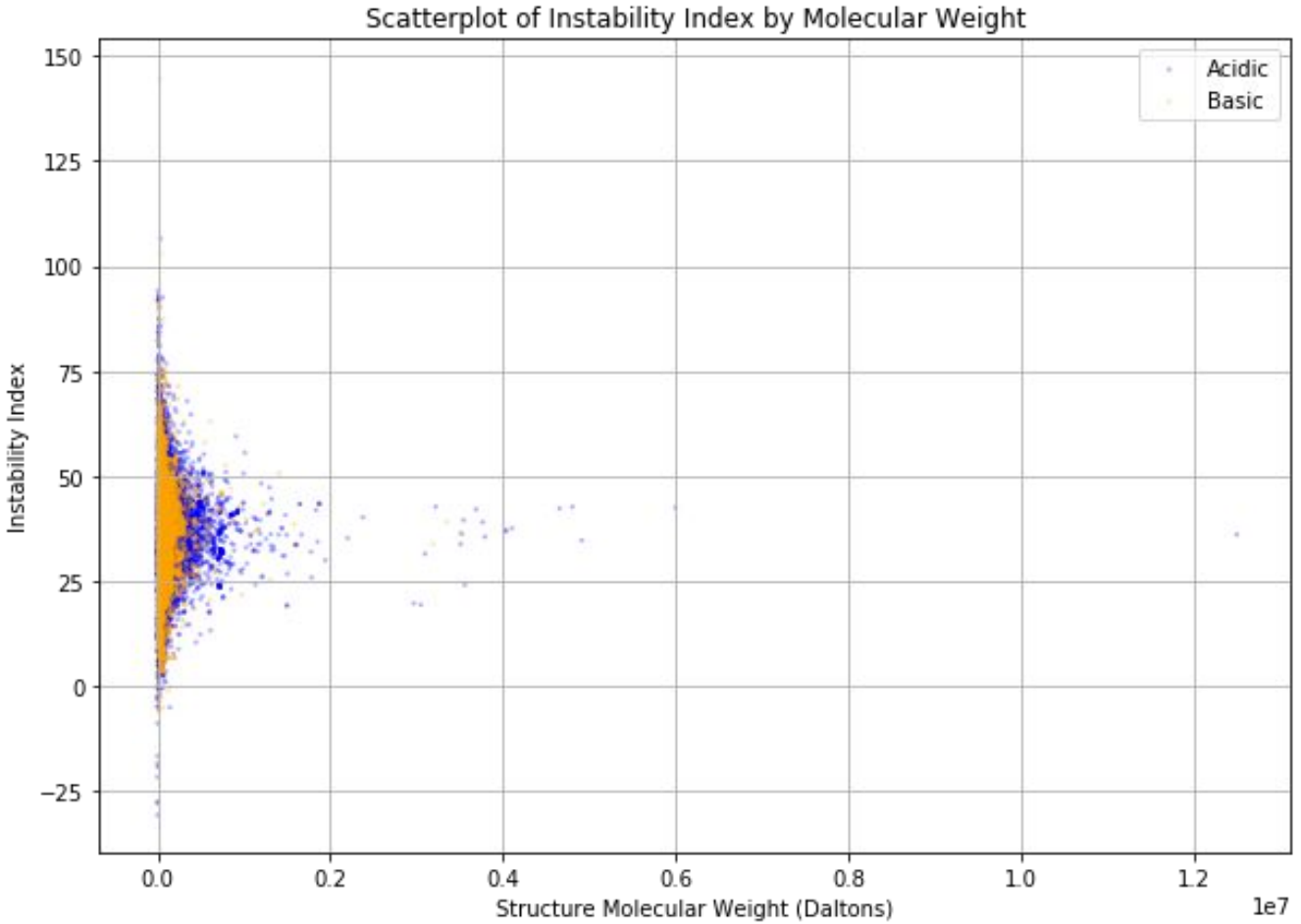
Could we create a classification model to predict whether a protein is an acid or a base given the protein's structure molecular weight in Daltons and its instability index?

Methods

I used supervised learning for my classification model. I used sklearn to test and train with the data from the protein database

Findings

After testing and training this data, I was able to build a classification model with 0.7454078499246272 7 accuracy. This shows that it is possible to build a model that can make a decent guess on protein acidity based on molecular weight and instability



Limitations

There are outliers and exceptions for this pattern. There are countless types of proteins, many of which we may not have discovered, that would prevent this classification model from being completely accurate.

Conclusions

There is a pattern with these features. However, the classification model built using just these attributes would not be reliable enough on its own for precise lab work or medicine (75% accuracy).

Acknowledgements

Protein database - David Dorner UCSD

References

Edx UCSD