Machine Learning Analysis of Protein Function Prediction

Matthew Segar
Tong Wu

Outline

Introduction

Methods

Results

Conclusion

Outline

Introduction

Methods

Results

Conclusion

A Dichotomy Exists

Sequencing has become increasingly easy

Databases (SwissProt) make it easy to upload data

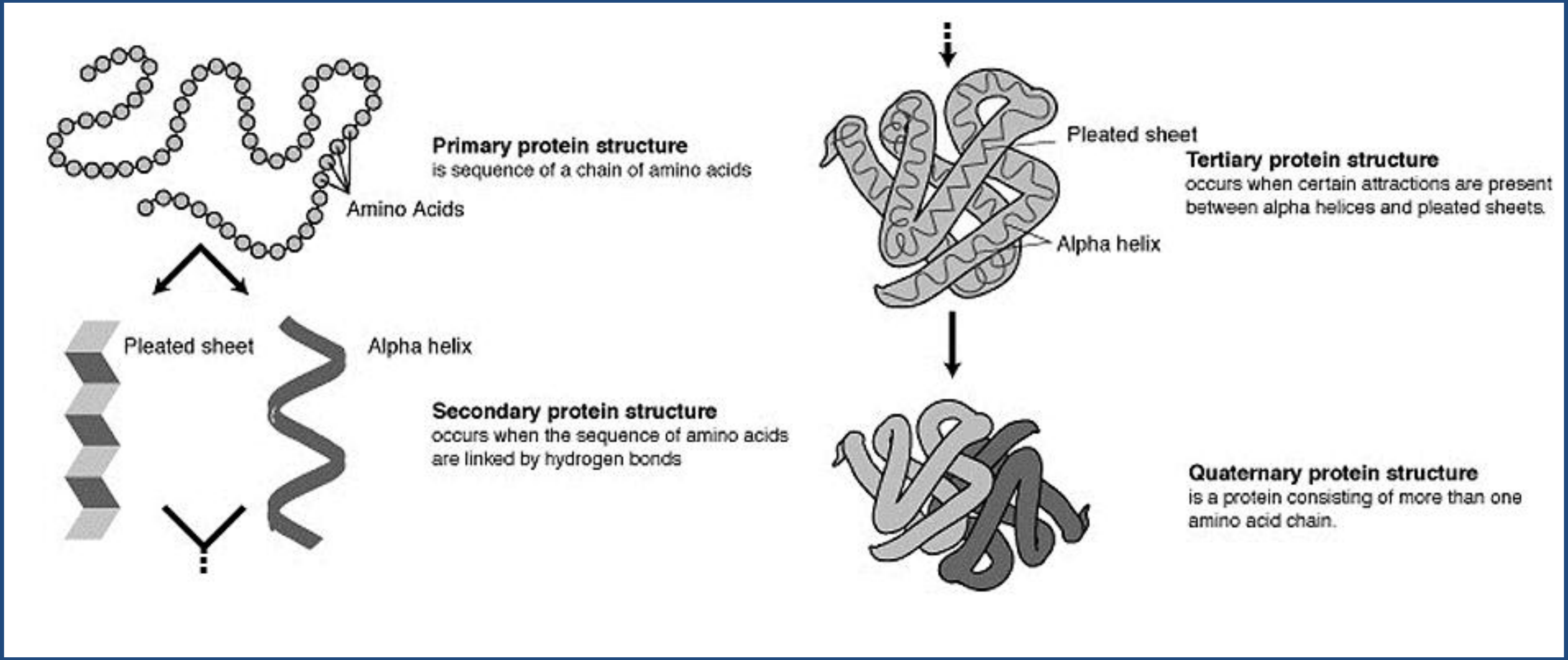
Data that is sequenced >>>> data that is annotated

Protein Function

“Anything and everything that happens to or through a protein” (Syed)

From sequence alone, the best predictors are only 90% accurate (Rost, Whisstock)

Protein Structure


Amino Acids
Primary protein structure
is sequence of a chain of amino acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Outline

Introduction

Methods

Results

Conclusion

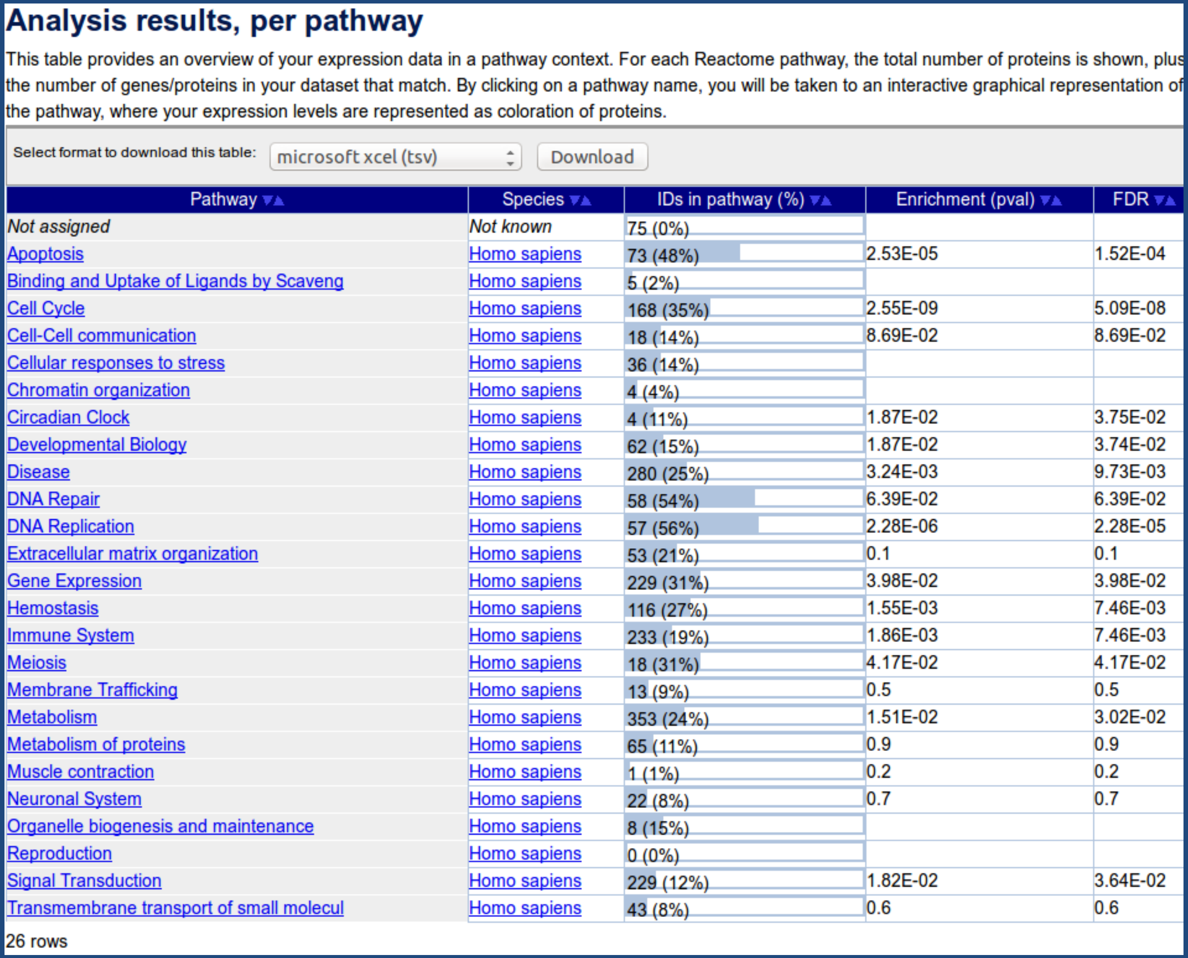
Data

Randomly selected 5000 protein sequences from SwissProt

Used *Reactome* to determine GO annotation

- Input list of Uniprot IDs
- Output broad function

Reactome



Analysis results, per pathway

This table provides an overview of your expression data in a pathway context. For each Reactome pathway, the total number of proteins is shown, plus the number of genes/proteins in your dataset that match. By clicking on a pathway name, you will be taken to an interactive graphical representation of the pathway, where your expression levels are represented as coloration of proteins.

Select format to download this table: microsoft xcel (tsv) Download

Pathway ▼▲	Species ▼▲	IDs in pathway (%) ▼▲	Enrichment (pval) ▼▲	FDR ▼▲
Not assigned	*Not known*	75 (0%)		
Apoptosis	Homo sapiens	73 (48%)	2.53E-05	1.52E-04
Binding and Uptake of Ligands by Scaveng	Homo sapiens	5 (2%)		
Cell Cycle	Homo sapiens	168 (35%)	2.55E-09	5.09E-08
Cell-Cell communication	Homo sapiens	18 (14%)	8.69E-02	8.69E-02
Cellular responses to stress	Homo sapiens	36 (14%)		
Chromatin organization	Homo sapiens	4 (4%)		
Circadian Clock	Homo sapiens	4 (11%)	1.87E-02	3.75E-02
Developmental Biology	Homo sapiens	62 (15%)	1.87E-02	3.74E-02
Disease	Homo sapiens	280 (25%)	3.24E-03	9.73E-03
DNA Repair	Homo sapiens	58 (54%)	6.39E-02	6.39E-02
DNA Replication	Homo sapiens	57 (56%)	2.28E-06	2.28E-05
Extracellular matrix organization	Homo sapiens	53 (21%)	0.1	0.1
Gene Expression	Homo sapiens	229 (31%)	3.98E-02	3.98E-02
Hemostasis	Homo sapiens	116 (27%)	1.55E-03	7.46E-03
Immune System	Homo sapiens	233 (19%)	1.86E-03	7.46E-03
Meiosis	Homo sapiens	18 (31%)	4.17E-02	4.17E-02
Membrane Trafficking	Homo sapiens	13 (9%)	0.5	0.5
Metabolism	Homo sapiens	353 (24%)	1.51E-02	3.02E-02
Metabolism of proteins	Homo sapiens	65 (11%)	0.9	0.9
Muscle contraction	Homo sapiens	1 (1%)	0.2	0.2
Neuronal System	Homo sapiens	22 (8%)	0.7	0.7
Organelle biogenesis and maintenance	Homo sapiens	8 (15%)		
Reproduction	Homo sapiens	0 (0%)		
Signal Transduction	Homo sapiens	229 (12%)	1.82E-02	3.64E-02
Transmembrane transport of small molecul	Homo sapiens	43 (8%)	0.6	0.6

26 rows

Filter

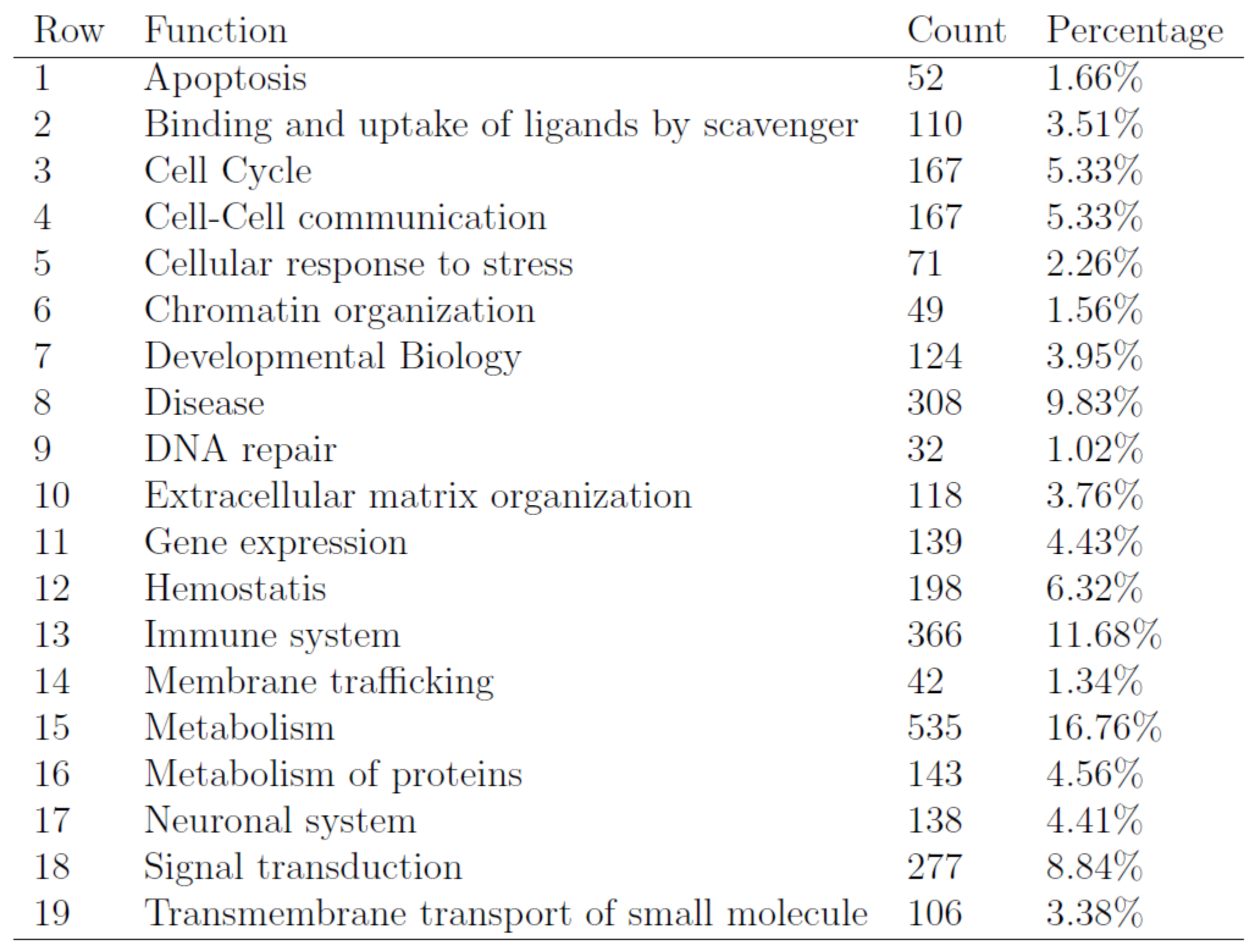
Only 3,085 protein sequences had functions

Removed functions with < 40 samples

Each function needed to constitute 1% of the total dataset

Resulted in 19 functions and 3,039 sequences

Data



Row	Function	Count	Percentage
1	Apoptosis	52	1.66%
2	Binding and uptake of ligands by scavenger	110	3.51%
3	Cell Cycle	167	5.33%
4	Cell-Cell communication	167	5.33%
5	Cellular response to stress	71	2.26%
6	Chromatin organization	49	1.56%
7	Developmental Biology	124	3.95%
8	Disease	308	9.83%
9	DNA repair	32	1.02%
10	Extracellular matrix organization	118	3.76%
11	Gene expression	139	4.43%
12	Hemostatis	198	6.32%
13	Immune system	366	11.68%
14	Membrane trafficking	42	1.34%
15	Metabolism	535	16.76%
16	Metabolism of proteins	143	4.56%
17	Neuronal system	138	4.41%
18	Signal transduction	277	8.84%
19	Transmembrane transport of small molecule	106	3.38%

Protein Sequence Features

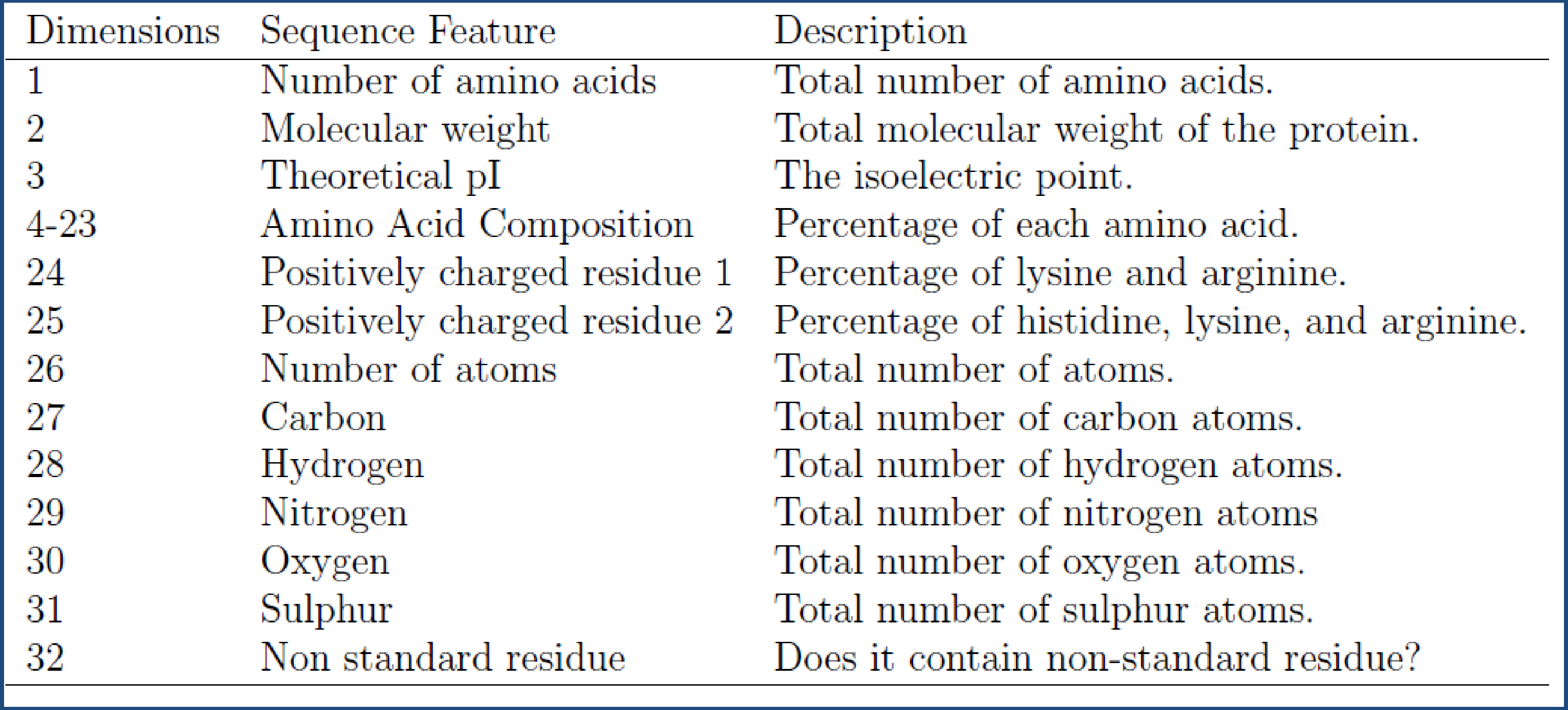
Extracted 32 features for each protein

Wrote custom C++ script to calculate most features

R was used for the rest

Output into ARFF format

Features

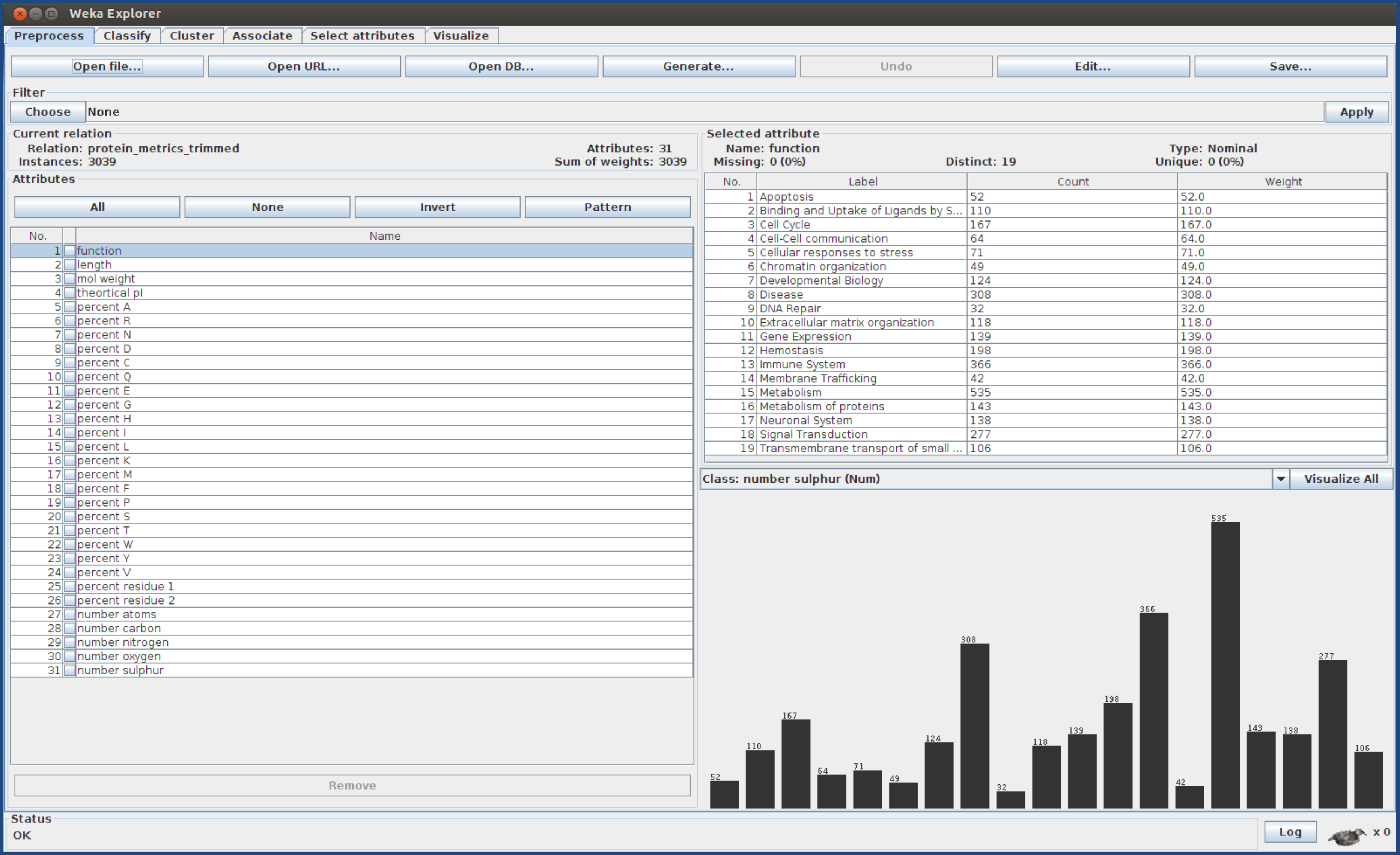


Dimensions	Sequence Feature	Description
1	Number of amino acids	Total number of amino acids.
2	Molecular weight	Total molecular weight of the protein.
3	Theoretical pI	The isoelectric point.
4-23	Amino Acid Composition	Percentage of each amino acid.
24	Positively charged residue 1	Percentage of lysine and arginine.
25	Positively charged residue 2	Percentage of histidine, lysine, and arginine.
26	Number of atoms	Total number of atoms.
27	Carbon	Total number of carbon atoms.
28	Hydrogen	Total number of hydrogen atoms.
29	Nitrogen	Total number of nitrogen atoms
30	Oxygen	Total number of oxygen atoms.
31	Sulphur	Total number of sulphur atoms.
32	Non standard residue	Does it contain non-standard residue?

ML Analysis

1. Decision Tree
2. Random Forest
3. Support Vector Machine
4. Neural Network

WEKA



Filter

None

Current relation

Relation: protein_metrics_trimmed
Instances: 3039
Attributes: 31
Sum of weights: 3039

Attributes

No.		Name
1	□	function
2	□	length
3	□	mol weight
4	□	theortical pI
5	□	percent A
6	□	percent R
7	□	percent N
8	□	percent D
9	□	percent C
10	□	percent Q
11	□	percent E
12	□	percent G
13	□	percent H
14	□	percent I
15	□	percent L
16	□	percent K
17	□	percent M
18	□	percent F
19	□	percent P
20	□	percent S
21	□	percent T
22	□	percent W
23	□	percent Y
24	□	percent V
25	□	percent residue 1
26	□	percent residue 2
27	□	number atoms
28	□	number carbon
29	□	number nitrogen
30	□	number oxygen
31	□	number sulphur

Selected attribute

Name: function
Missing: 0 (0%)
Distinct: 19
Type: Nominal
Unique: 0 (0%)

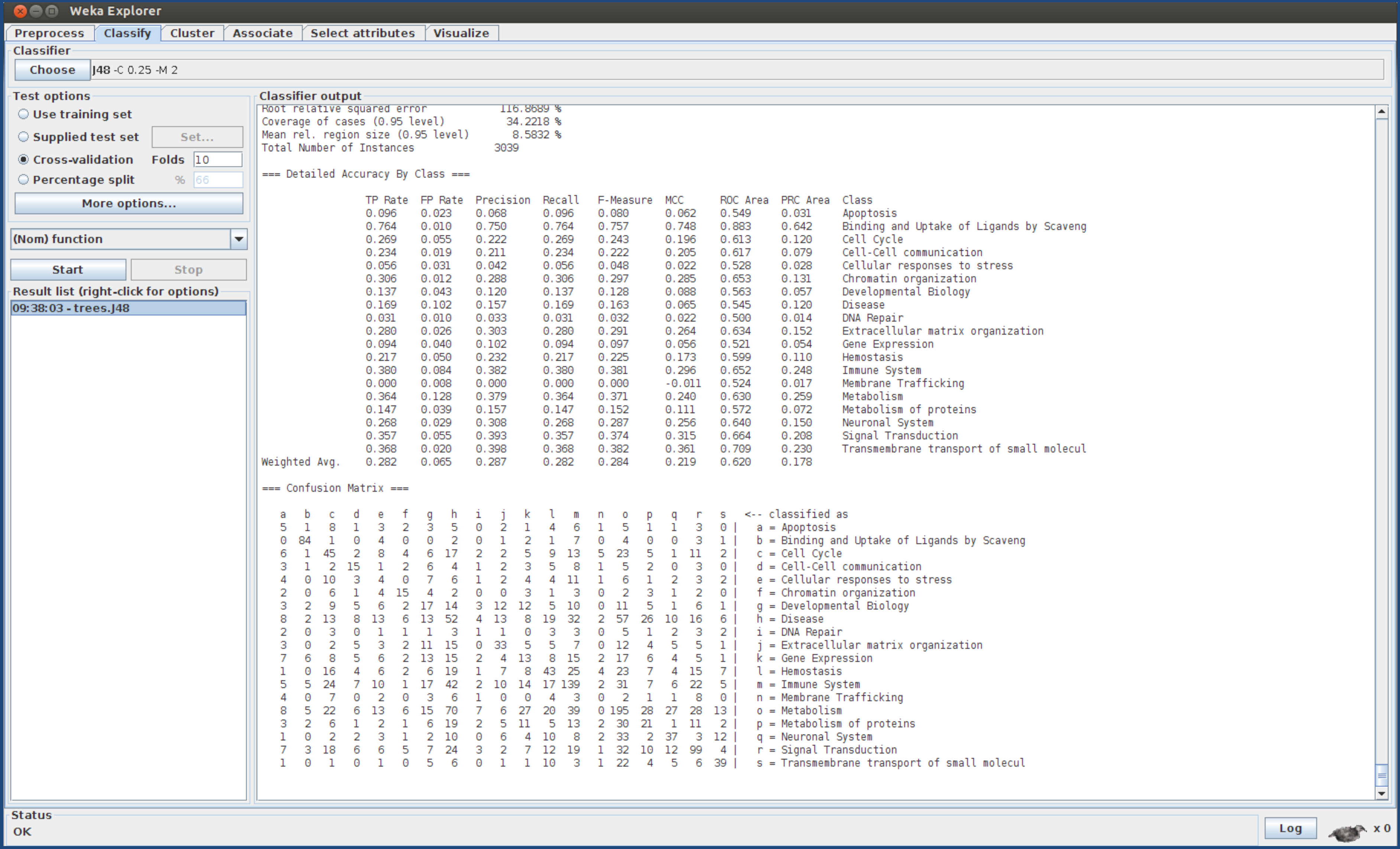
No.	Label	Count	Weight
1	Apoptosis	52	52.0
2	Binding and Uptake of Ligands by S...	110	110.0
3	Cell Cycle	167	167.0
4	Cell-Cell communication	64	64.0
5	Cellular responses to stress	71	71.0
6	Chromatin organization	49	49.0
7	Developmental Biology	124	124.0
8	Disease	308	308.0
9	DNA Repair	32	32.0
10	Extracellular matrix organization	118	118.0
11	Gene Expression	139	139.0
12	Hemostasis	198	198.0
13	Immune System	366	366.0
14	Membrane Trafficking	42	42.0
15	Metabolism	535	535.0
16	Metabolism of proteins	143	143.0
17	Neuronal System	138	138.0
18	Signal Transduction	277	277.0
19	Transmembrane transport of small ...	106	106.0

Class: number sulphur (Num)

Status

OK

WEKA



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) function

Result list (right-click for options)

09:38:03 - trees.J48

Classifier output

```
Root relative squared error            116.8689 %
Coverage of cases (0.95 level)          34.2218 %
Mean rel. region size (0.95 level)       8.5832 %
Total Number of Instances             3039

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.096    0.023    0.068      0.096    0.080      0.062    0.549     0.031     Apoptosis
                 0.764    0.010    0.750      0.764    0.757      0.748    0.883     0.642     Binding and Uptake of Ligands by Scaveng
                 0.269    0.055    0.222      0.269    0.243      0.196    0.613     0.120     Cell Cycle
                 0.234    0.019    0.211      0.234    0.222      0.205    0.617     0.079     Cell-Cell communication
                 0.056    0.031    0.042      0.056    0.048      0.022    0.528     0.028     Cellular responses to stress
                 0.306    0.012    0.288      0.306    0.297      0.285    0.653     0.131     Chromatin organization
                 0.137    0.043    0.120      0.137    0.128      0.088    0.563     0.057     Developmental Biology
                 0.169    0.102    0.157      0.169    0.163      0.065    0.545     0.120     Disease
                 0.031    0.010    0.033      0.031    0.032      0.022    0.500     0.014     DNA Repair
                 0.280    0.026    0.303      0.280    0.291      0.264    0.634     0.152     Extracellular matrix organization
                 0.094    0.040    0.102      0.094    0.097      0.056    0.521     0.054     Gene Expression
                 0.217    0.050    0.232      0.217    0.225      0.173    0.599     0.110     Hemostasis
                 0.380    0.084    0.382      0.380    0.381      0.296    0.652     0.248     Immune System
                 0.000    0.008    0.000      0.000    0.000      -0.011   0.524     0.017     Membrane Trafficking
                 0.364    0.128    0.379      0.364    0.371      0.240    0.630     0.259     Metabolism
                 0.147    0.039    0.157      0.147    0.152      0.111    0.572     0.072     Metabolism of proteins
                 0.268    0.029    0.308      0.268    0.287      0.256    0.640     0.150     Neuronal System
                 0.357    0.055    0.393      0.357    0.374      0.315    0.664     0.208     Signal Transduction
                 0.368    0.020    0.398      0.368    0.382      0.361    0.709     0.230     Transmembrane transport of small molecul
Weighted Avg.    0.282    0.065    0.287      0.282    0.284      0.219    0.620     0.178     

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   s   <-- classified as
   5   1   8   1   3   2   3   5   0   2   1   4   6   1   5   1   1   3   0 |   a = Apoptosis
   0  84   1   0   4   0   0   2   0   1   2   1   7   0   4   0   0   3   1 |   b = Binding and Uptake of Ligands by Scaveng
   6   1  45   2   8   4   6  17   2   2   5   9  13   5  23   5   1  11   2 |   c = Cell Cycle
   3   1   2  15   1   2   6   4   1   2   3   5   8   1   5   2   0   3   0 |   d = Cell-Cell communication
   4   0  10   3   4   0   7   6   1   2   4   4  11   1   6   1   2   3   2 |   e = Cellular responses to stress
   2   0   6   1   4  15   4   2   0   0   3   1   3   0   2   3   1   2   0 |   f = Chromatin organization
   3   2   9   5   6   2  17  14   3  12  12   5  10   0  11   5   1   6   1 |   g = Developmental Biology
   8   2  13   8  13   6  13  52   4  13   8  19  32   2  57  26  10  16   6 |   h = Disease
   2   0   3   0   1   1   1   3   1   1   0   3   3   0   5   1   2   3   2 |   i = DNA Repair
   3   0   2   5   3   2  11  15   0  33   5   5   7   0  12   4   5   5   1 |   j = Extracellular matrix organization
   7   6   8   5   6   2  13  15   2   4  13   8  15   2  17   6   4   5   1 |   k = Gene Expression
   1   0  16   4   6   2   6  19   1   7   8  43  25   4  23   7   4  15   7 |   l = Hemostasis
   5   5  24   7  10   1  17  42   2  10  14  17 139   2  31   7   6  22   5 |   m = Immune System
   4   0   7   0   2   0   3   6   1   0   0   4   3   0   2   1   1   8   0 |   n = Membrane Trafficking
   8   5  22   6  13   6  15  70   7   6  27  20  39   0 195  28  27  28  13 |   o = Metabolism
   3   2   6   1   2   1   6  19   2   5  11   5  13   2  30  21   1  11   2 |   p = Metabolism of proteins
   1   0   2   2   3   1   2  10   0   6   4  10   8   2  33   2  37   3  12 |   q = Neuronal System
   7   3  18   6   6   5   7  24   3   2   7  12  19   1  32  10  12  99   4 |   r = Signal Transduction
   1   0   1   0   1   0   5   6   0   1   1  10   3   1  22   4   5   6  39 |   s = Transmembrane transport of small molecul
```

Status

OK

Outline

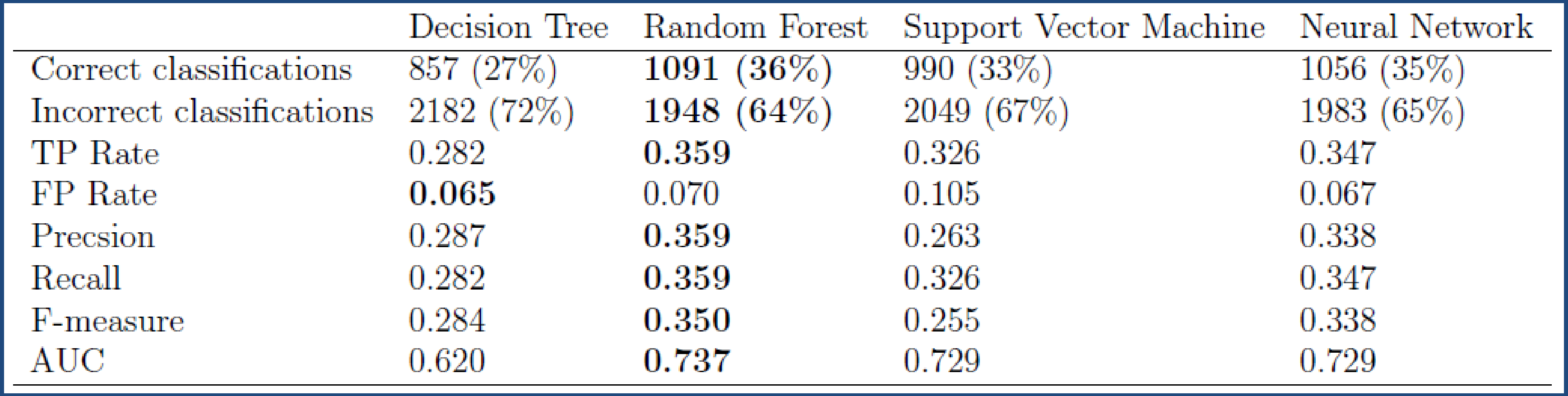
Introduction

Methods

Results

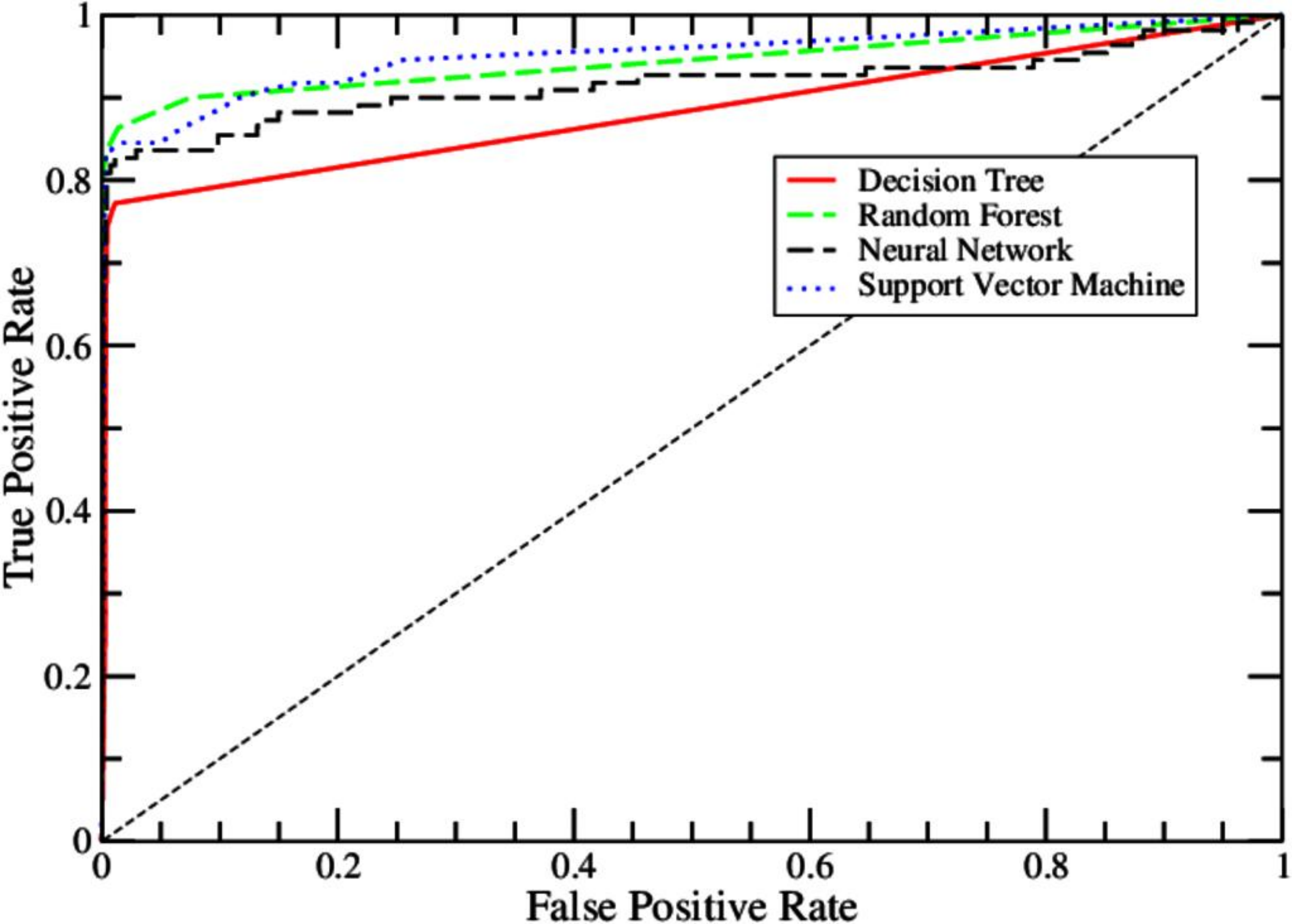
Conclusion

Results

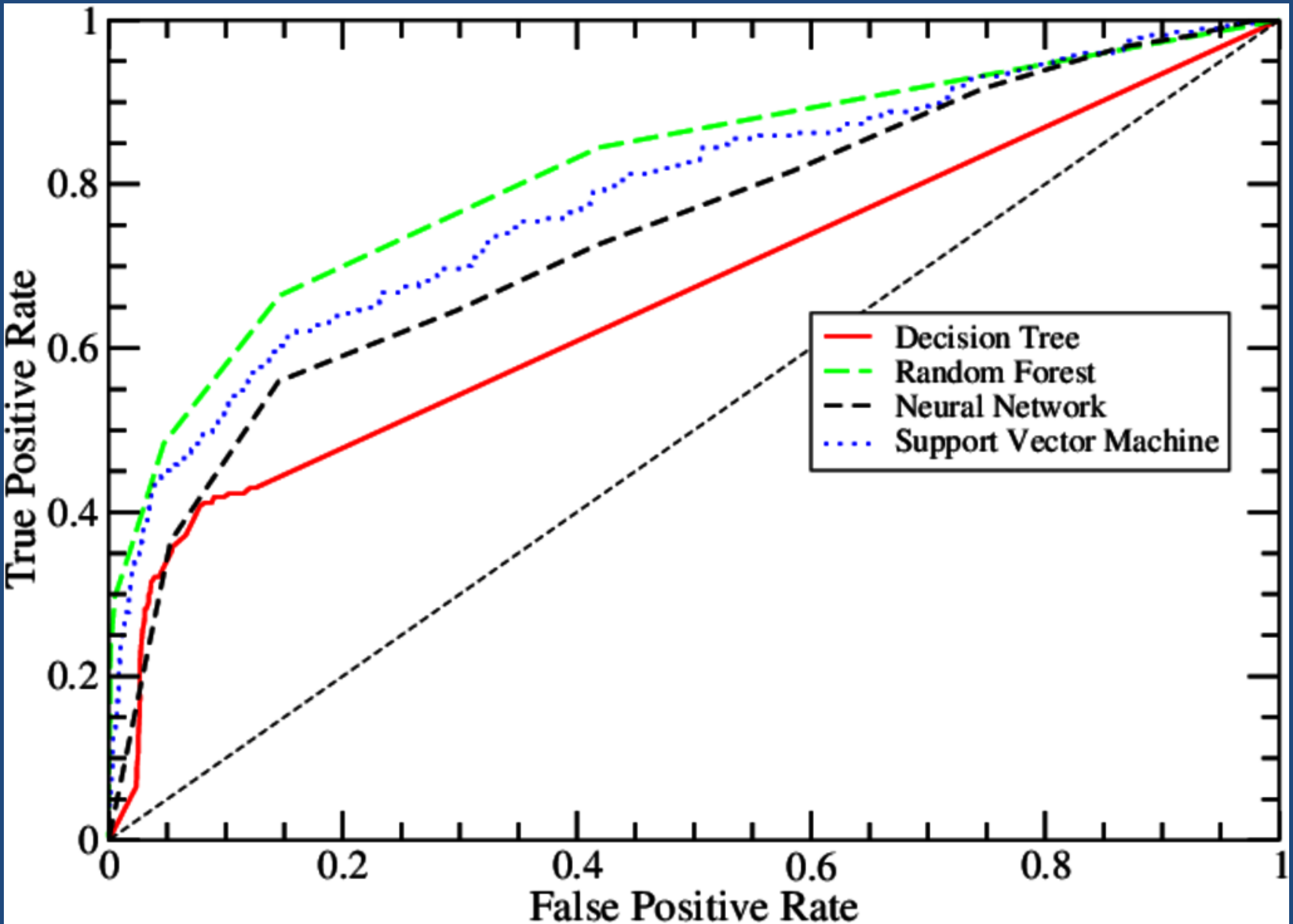


	Decision Tree	Random Forest	Support Vector Machine	Neural Network
Correct classifications	857 (27%)	**1091 (36%)**	990 (33%)	1056 (35%)
Incorrect classifications	2182 (72%)	**1948 (64%)**	2049 (67%)	1983 (65%)
TP Rate	0.282	**0.359**	0.326	0.347
FP Rate	**0.065**	0.070	0.105	0.067
Precsion	0.287	**0.359**	0.263	0.338
Recall	0.282	**0.359**	0.326	0.347
F-measure	0.284	**0.350**	0.255	0.338
AUC	0.620	**0.737**	0.729	0.729

ROC – Binding and Uptake



ROC – Signal Transduction



Outline

Introduction

Methods

Results

Conclusion

Conclusion

Random Forest performed best

Best ML application only had 36% correctness

AUC was over 70%

Improvements

Need more data (upwards of 10,000)

More even distribution of functions

Need more protein characteristics

GitHub

Available online at:

www.github.com/msegar/IT529-Protein-Prediction