Prediction of host specificity based on PB1

Petter Byström and Clàudia González

Outline

1. Feature selection - mcfs
2. Creating the model - Identification of rules with Rosetta
3. Evaluation
4. Visualization

Preprossesing

Split data set into training (70%) and test (30%)

Remove no data, treat it as discrete

Feature selection

MCFS

result <- mcfs(Host ~ ., data, projections = 1500, projectionSize = 0.1, splits = 5, splitSetSize = 500, cutoffPermutations = 6, threadsNumber = 8)

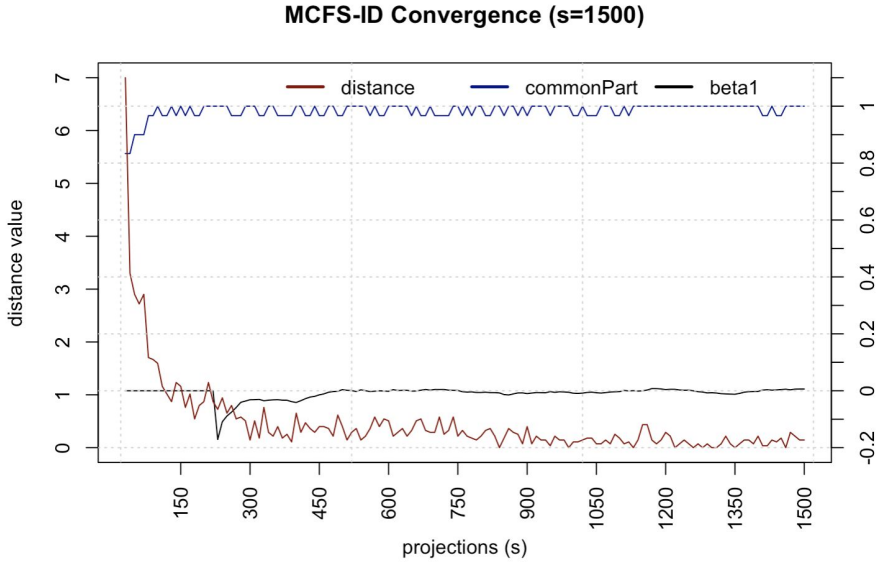
36 features found

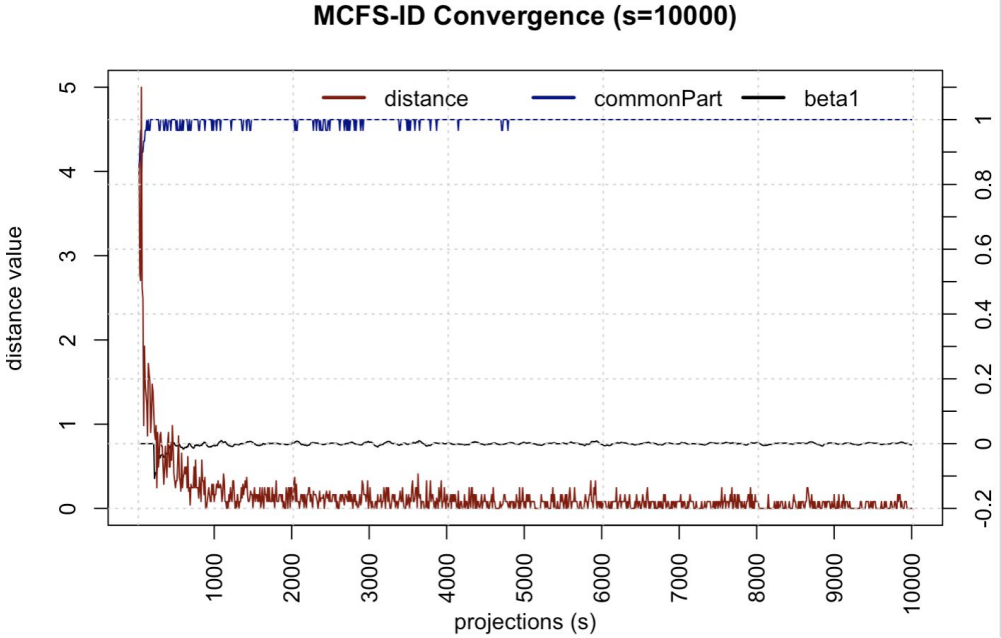
Evaluation of MCFs

```
Accuracy = 91.30%
WeightedAccuracy = 92.85%
```

Improvements: cutoffPermuations = 20, more projections

Feature selection





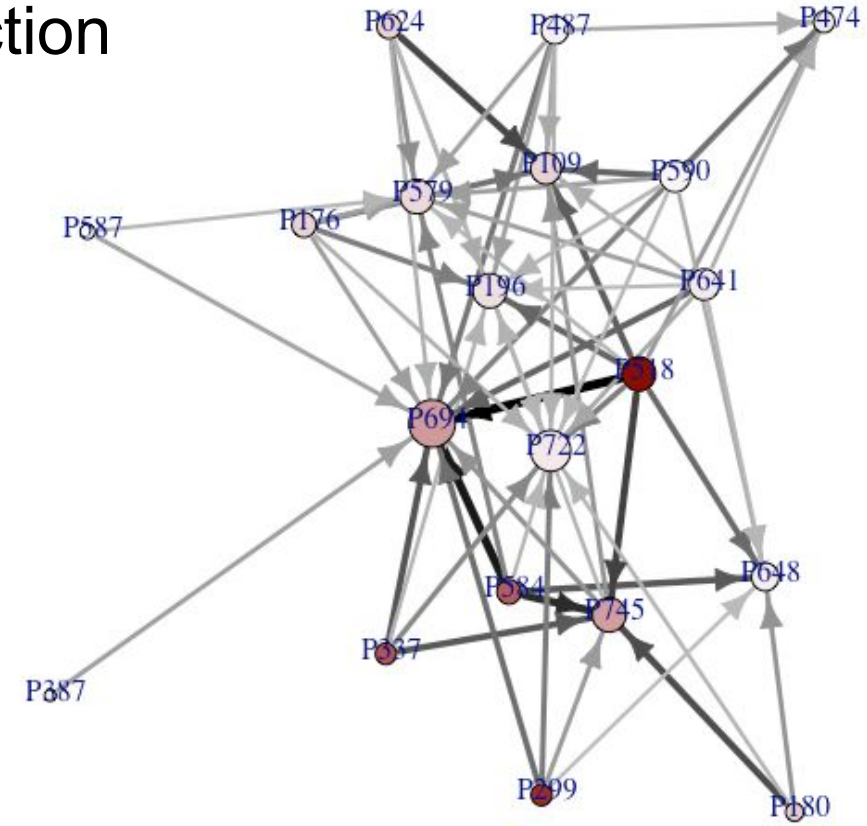
Feature selection

```
> Sig_result[1:20,]
    position attribute projections classifiers     nodes        RI
518        1      P518         153   0.9516340 0.9516340 0.6396244
299        2      P299         149   0.8979866 0.8979866 0.5575160
337        3      P337         140   0.8071428 0.8071428 0.4783977
582        4      P584         147   0.7455782 0.7455782 0.4351913
692        5      P694         146   0.8684931 0.8739726 0.3462521
742        6      P745         150   0.8786666 0.9186667 0.3340654
176        7      P176         141   0.6567376 0.6652482 0.2449936
180        8      P180         148   0.6000000 0.6000000 0.2337762
622        9      P624         181   0.6364641 0.6430939 0.2334032
109       10      P109         152   0.6486842 0.6500000 0.2187381
196       11      P196         145   0.5655172 0.5820690 0.2118568
639       12      P641         147   0.5564626 0.5564626 0.2036305
387       13      P387         157   0.5184714 0.5197452 0.2008733
577       14      P579         160   0.6287500 0.6387500 0.1937117
720       15      P722         154   0.6792208 0.6792208 0.1780564
646       16      P648         143   0.6321678 0.6391608 0.1757603
474       17      P474         145   0.5627586 0.5793104 0.1671090
588       18      P590         167   0.4514970 0.4586826 0.1526724
487       19      P487         129   0.4108527 0.4108527 0.1503529
211       20      P211         153   0.4000000 0.4000000 0.1491660
```

Most important features

nodes	RI			
P518	153	0.9516340	0.9516340	0.6396244
P299	149	0.8979866	0.8979866	0.5575160
P337	140	0.8071428	0.8071428	0.4783977
P584	147	0.7455782	0.7455782	0.4351913
P694	146	0.8684931	0.8739726	0.3462521

Feature selection



Creating the model

Rosetta uses our most significant attributes as input

```
data <- rosetta(table_significant, roc = TRUE, clroc = "Human", discrete = T)
```

Reduction method: Johnson and Genetic

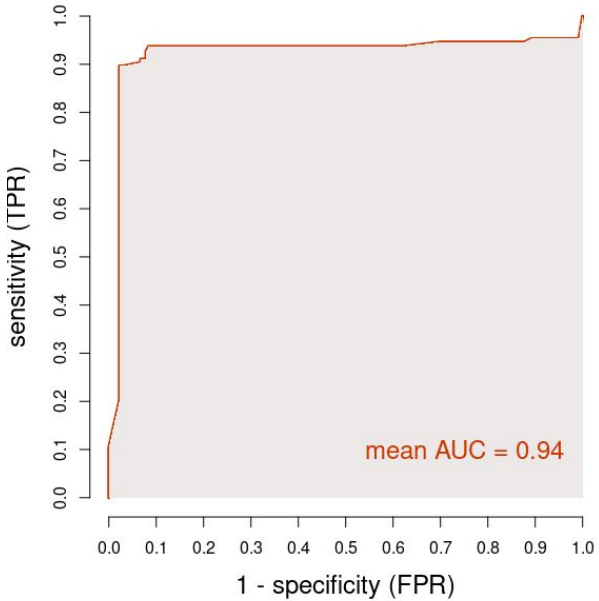
Top 5 most significant rules Johnson

Rule	Length	Acc	Support	P-value
IF P299(L) THEN Avian	1	0.96429	108	2.334259e-41
IF P337(V) THEN Avian	1	0.96076	107	1.563741e-40
IF P584(E) THEN Avian	1	0.96640	104	1.587791e-39
IF P176(N) THEN Human	1	0.96649	73	1.255420e-29
IF P362(R) THEN Human	1	0.96382	71	1.085451e-25

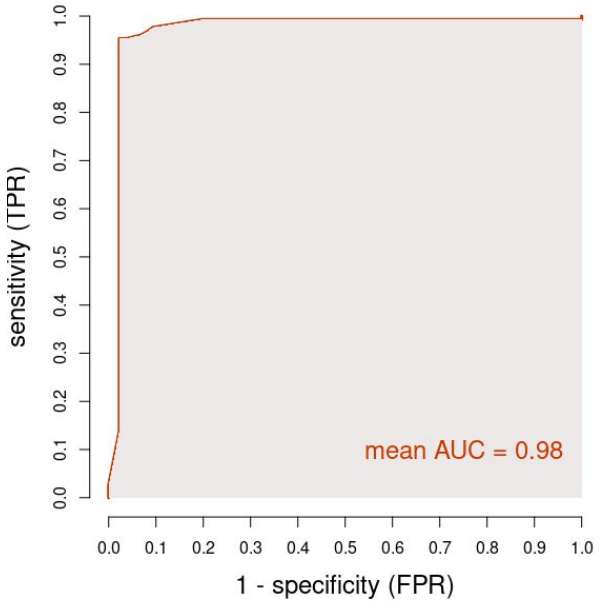
Top 5 most significant rules Genetic

Rule	Length	Acc	Support	P-value
IF P518(V) AND P722(V) THEN Human	2	0.98941	94	5.798414e-45
IF P299(I) THEN Human	1	0.97980	97	3.315916e-44
IF P387(K) AND P518(V) AND P722(V) THEN Human	3	1.00000	91	2.531160e-42
IF P337(I) AND P387(K) AND P722(V) THEN Human	3	0.98936	93	3.790874e-42
IF P518(V) AND P584(D) AND P722(V) THEN Human	3	1.00000	90	1.759759e-41

ROC johnson vs genetic



Johnson



Genetic

Evaluation

Evaluate rules from rosetta to use rules on the test data

```
predcitClass(Test,rules_h, discrete =True,normalize=True,normalizeMethod="rss",validate=TRUE,decision)
```

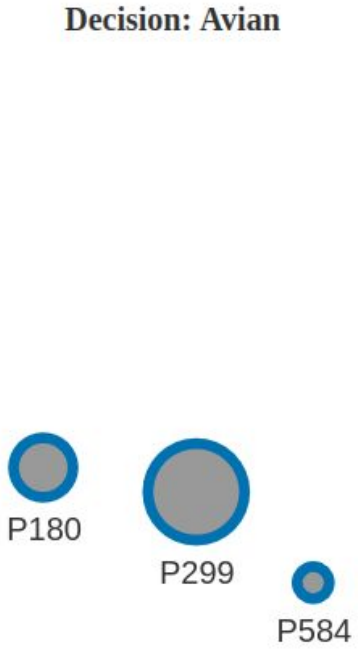
	Avian	Human	currentClass	predictedClass
8	0.000000000	0.024767802	Human	Human
12	0.000000000	0.000000000	Human	Avian
14	0.000000000	0.024767802	Human	Human
30	0.000000000	0.024767802	Human	Human
34	0.000000000	0.024767802	Human	Human

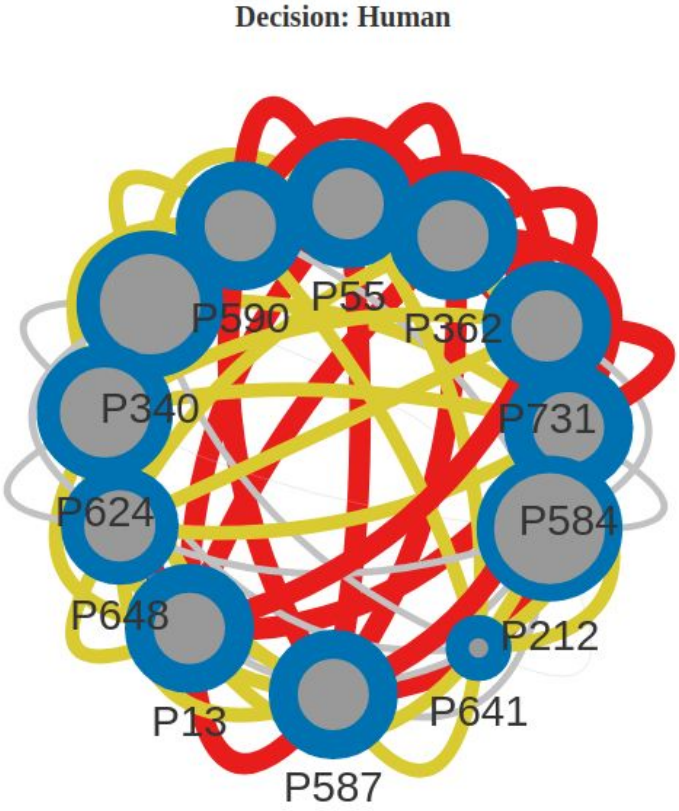
Johnson
Accuracy : 0.9388

	Avian	Human	currentClass	predictedClass
1	0.000000e+00	1.437335e-02	Human	Human
2	0.000000e+00	1.437335e-02	Human	Human
3	0.000000e+00	1.437335e-02	Human	Human
4	0.000000e+00	1.437335e-02	Human	Human
5	0.000000e+00	1.437335e-02	Human	Human

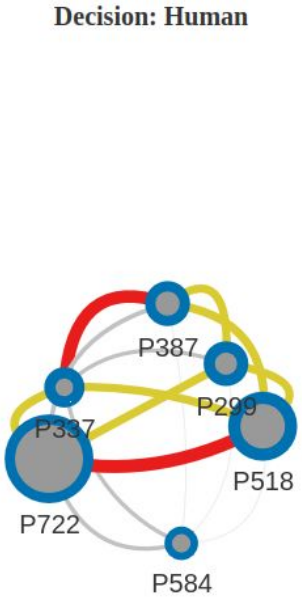
Genetic
Accuracy : 0.9622

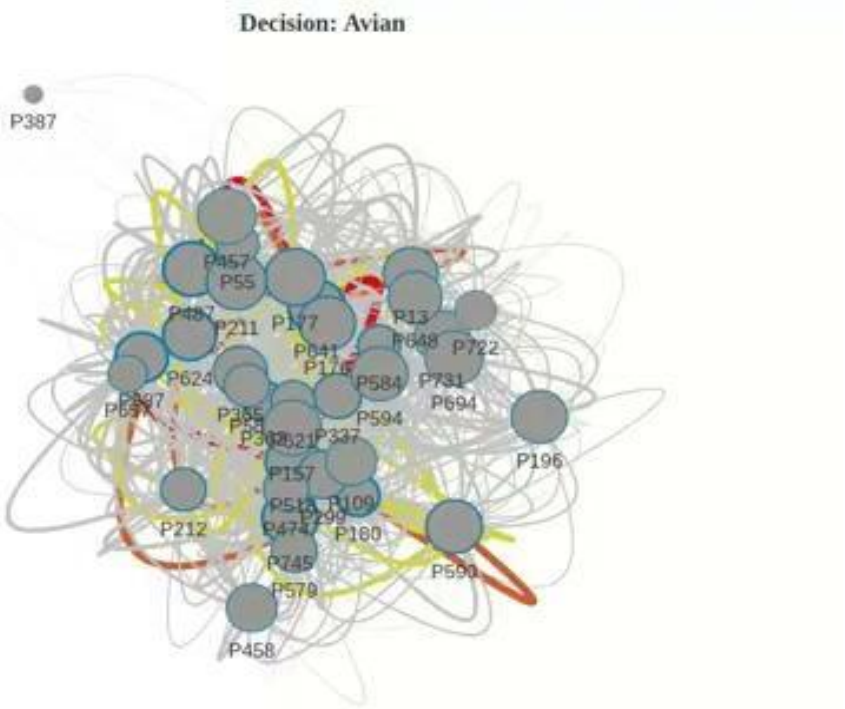
Visualisation of rules - Johnson





Visualisation of rules - Genetic





Conclusion

Most important features: P518 P299 P337 P584

Most important rules: `IF P299(L) THEN Avian, IF P337(V) THEN Avian, IF P584(E) THEN Avian`

Accuracy: `0.9388`