

Experiment 1 : Predicting the family of a test sequence based on minimum compression

Dataset : **Astral SCOPE 2.05** genetic domain sequence subsets, based on PDB SEQRES records, with less than 95% identity to each other

Observation

- I) 37% of families have only single member (1751/4694)
- II) 67% of families have less than 4 members

Super families chosen for experiment

- d.122.1 (has 5 families with 8, 8, 10, 15, 6 members in order)
- b.47.1 (has 5 families with 14, 88, 19, 33, 20 members in order)
- b.1.1 (has 5 families with 11, 161, 810, 639, 89 members in order)

Test dataset : 2 members from each family picked randomly

Each test case is added to each of the family corpora and segmented. The family which achieves maximum gain in Description Length (maximum compression) is the predicted class label.

d.122.1 (Precision is 80.0)

test case 1

family d.122.1.0 compression 34.8940717028
family d.122.1.1 compression 250.068618694
family d.122.1.2 compression 0
family d.122.1.3 compression 0
family d.122.1.4 compression 3.75199770695
Predicted : d.122.1.1 Actual : d.122.1.1
True

test case 2

family d.122.1.0 compression 27.0062003023
family d.122.1.1 compression 265.882666579
family d.122.1.2 compression 0
family d.122.1.3 compression 0
family d.122.1.4 compression 0.839353235968
Predicted : d.122.1.1 Actual : d.122.1.1
True

test case 3

family d.122.1.0 compression 4.93310525314
family d.122.1.1 compression 4.96105808329
family d.122.1.2 compression 7.75151461379
family d.122.1.3 compression 2.20198104654

family d.122.1.4 compression 2.77426265724
Predicted : d.122.1.2 Actual : d.122.1.2
True

test case 4

family d.122.1.0 compression 104.718652739
family d.122.1.1 compression 9.10234433406
family d.122.1.2 compression 24.4185407832
family d.122.1.3 compression 1.42369207153
family d.122.1.4 compression 0.282842028293
Predicted : d.122.1.0 Actual : d.122.1.2
False

test case 5

family d.122.1.0 compression 2.803745609
family d.122.1.1 compression 0.977875727084
family d.122.1.2 compression 0.0263043924611
family d.122.1.3 compression 1.5531548365
family d.122.1.4 compression 0
Predicted : d.122.1.0 Actual : d.122.1.3
False

test case 6

family d.122.1.0 compression 2.41840438549
family d.122.1.1 compression 5.71134664813
family d.122.1.2 compression 0
family d.122.1.3 compression 148.343211833
family d.122.1.4 compression 0
Predicted : d.122.1.3 Actual : d.122.1.3
True

test case 7

family d.122.1.0 compression 12.0459868775
family d.122.1.1 compression 1.39685398876
family d.122.1.2 compression 0.386547040209
family d.122.1.3 compression 0
family d.122.1.4 compression 19.2243675355
Predicted : d.122.1.4 Actual : d.122.1.4
True

test case 8

family d.122.1.0 compression 104.268019723
family d.122.1.1 compression 1.94493735123
family d.122.1.2 compression 0
family d.122.1.3 compression 0
family d.122.1.4 compression 442.470649239
Predicted : d.122.1.4 Actual : d.122.1.4
True

test case 9

family d.122.1.0 compression 162.66168476
family d.122.1.1 compression 10.3002261584
family d.122.1.2 compression 17.9617394413
family d.122.1.3 compression 0
family d.122.1.4 compression 3.09998620696
Predicted : d.122.1.0 Actual : d.122.1.0
True

test case 10

family d.122.1.0 compression 113.127280649
family d.122.1.1 compression 4.92654665543
family d.122.1.2 compression 56.7771619326
family d.122.1.3 compression 0
family d.122.1.4 compression 0.857349303993
Predicted : d.122.1.0 Actual : d.122.1.0
True

b.47.1 (Precision : 50%)

test case 1

family b.47.1.0 compression 1.27921427464
family b.47.1.1 compression 8.77749411104
family b.47.1.2 compression 9.41560858653
family b.47.1.3 compression 1.43099098207
family b.47.1.4 compression 0.0370671809674
Predicted : b.47.1.2 Actual : b.47.1.1
False

test case 2

family b.47.1.0 compression 48.5222164585
family b.47.1.1 compression 13.4926772862
family b.47.1.2 compression 13.17290307
family b.47.1.3 compression 7.20569778787
family b.47.1.4 compression 0
Predicted : b.47.1.0 Actual : b.47.1.1
False

test case 3

family b.47.1.0 compression 25.7460361008
family b.47.1.1 compression 21.0499622293
family b.47.1.2 compression 66.4150836715
family b.47.1.3 compression 6.76403843142
family b.47.1.4 compression 0.311869039519
Predicted : b.47.1.2 Actual : b.47.1.2
True

test case 4

family b.47.1.0 compression 33.0809343755

family b.47.1.1 compression 14.7666716716

family b.47.1.2 compression 166.926670749

family b.47.1.3 compression 5.36494482941

family b.47.1.4 compression 5.75430207654

Predicted : b.47.1.2 Actual : b.47.1.2

True

test case 5

family b.47.1.0 compression 2.69172854435

family b.47.1.1 compression 2.02484625186

family b.47.1.2 compression 8.76016522603

family b.47.1.3 compression 38.5877114174

family b.47.1.4 compression 0

Predicted : b.47.1.3 Actual : b.47.1.3

True

test case 6

family b.47.1.0 compression 2.73732586472

family b.47.1.1 compression 2.20383480755

family b.47.1.2 compression 8.01896961223

family b.47.1.3 compression 77.7746340366

family b.47.1.4 compression 0

Predicted : b.47.1.3 Actual : b.47.1.3

True

test case 7

family b.47.1.0 compression 0.16687901655

family b.47.1.1 compression 0

family b.47.1.2 compression 3.23537879808

family b.47.1.3 compression 10.7815620734

family b.47.1.4 compression 79.8350521233

Predicted : b.47.1.4 Actual : b.47.1.4

True

test case 8

family b.47.1.0 compression 32.7660588246

family b.47.1.1 compression 0.0106812567892

family b.47.1.2 compression 8.31142714694

family b.47.1.3 compression 5.42044075559

family b.47.1.4 compression 32.5489588618

Predicted : b.47.1.0 Actual : b.47.1.4

False

test case 9

family b.47.1.0 compression 29.4090211971

family b.47.1.1 compression 20.331547621

family b.47.1.2 compression 94.8212942778

family b.47.1.3 compression 1.47111477853
family b.47.1.4 compression 1.3102302389
Predicted : b.47.1.2 Actual : b.47.1.0
False

test case 10
family b.47.1.0 compression 29.2698853489
family b.47.1.1 compression 17.9979397762
family b.47.1.2 compression 70.9064576175
family b.47.1.3 compression 11.7922418627
family b.47.1.4 compression 0.612212437791
Predicted : b.47.1.2 Actual : b.47.1.0
False

b.1.1 (Precision : 50 %)

test case 1
family b.1.1.0 compression 67.1543570617
family b.1.1.1 compression 201.620459817
family b.1.1.2 compression 4.91180780494
family b.1.1.3 compression 0
family b.1.1.4 compression 0.234978621938
Predicted : b.1.1.1 Actual : b.1.1.1
True

test case 2
family b.1.1.0 compression 90.8544462922
family b.1.1.1 compression 261.200478176
family b.1.1.2 compression 7.89913246742
family b.1.1.3 compression 0
family b.1.1.4 compression 0
Predicted : b.1.1.1 Actual : b.1.1.1
True

test case 3
family b.1.1.0 compression 8.52899916653
family b.1.1.1 compression 5.82779398162
family b.1.1.2 compression 267.798556069
family b.1.1.3 compression 0
family b.1.1.4 compression 4.51692628658
Predicted : b.1.1.2 Actual : b.1.1.2
True

test case 4
family b.1.1.0 compression 4.75322530072
family b.1.1.1 compression 5.06672575284

family b.1.1.2 compression 198.677973193
family b.1.1.3 compression 0
family b.1.1.4 compression 0.637083870224
Predicted : b.1.1.2 Actual : b.1.1.2
True

test case 5
family b.1.1.0 compression 5.03146524456
family b.1.1.1 compression 12.0727212935
family b.1.1.2 compression 3.16362218423
family b.1.1.3 compression 3.64367646274
family b.1.1.4 compression 2.79971419048
Predicted : b.1.1.1 Actual : b.1.1.3
False

test case 6
family b.1.1.0 compression 7.4835412202
family b.1.1.1 compression 12.1127408473
family b.1.1.2 compression 1.85587203089
family b.1.1.3 compression 2.83397490719
family b.1.1.4 compression 2.81059812104
Predicted : b.1.1.1 Actual : b.1.1.3
False

test case 7
family b.1.1.0 compression 13.877957567
family b.1.1.1 compression 13.2250867732
family b.1.1.2 compression 6.5548911103
family b.1.1.3 compression 0
family b.1.1.4 compression 0
Predicted : b.1.1.0 Actual : b.1.1.4
False

test case 8
family b.1.1.0 compression 17.0483803239
family b.1.1.1 compression 16.2338462145
family b.1.1.2 compression 17.409726409
family b.1.1.3 compression 0
family b.1.1.4 compression 3.35362425695
Predicted : b.1.1.2 Actual : b.1.1.4
False

test case 9
family b.1.1.0 compression 142.693184805
family b.1.1.1 compression 117.769992839
family b.1.1.2 compression 8.15272823552
family b.1.1.3 compression 0
family b.1.1.4 compression 1.68388977442

Predicted : b.1.1.0 Actual : b.1.1.0
True

test case 10

family b.1.1.0 compression 8.48786607497

family b.1.1.1 compression 10.2009988499

family b.1.1.2 compression 18.6041336895

family b.1.1.3 compression 0

family b.1.1.4 compression 0

Predicted : b.1.1.2 Actual : b.1.1.0

False
