**Analysis of SVM performance with variations in input feature set**

In our current solution to domain identification problem, we are employing SVM at two stages, first to classify between single & multi-domain proteins and next in classifying between multi-domain proteins. The below set of experiments compare the performance of SVM at each stage when different training/testing input feature sets are used. The test is done for benchmark\_2, benchmark\_3, ASTRAL SCOP30 and our own created dataset.

Please note that the features are as follows: L=>Length, I=>Interaction Strength, R=>Radius Of Gyration & D=>Density.

For Phase I SVM: Classifying single & multi-domain proteins

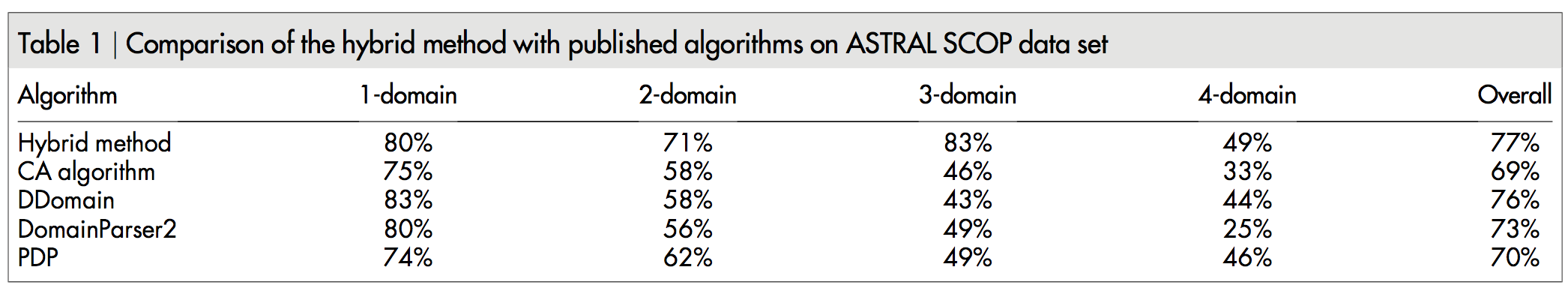
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature Set** | **Benchmark2** | | | | | | **Benchmark3** | | | | | | | | **ASTRAL SCOP30** | | | | | | | | | **Self Created** | | | | | | | | | | | | |
|  | Single Correct | Multi Correct | | | Overall | Single Correct | | Multi Correct | | | | Overall | | Single Correct | | | Multi Correct | | | | | Overall | | | Single Correct | | | Multi Correct | | | | | | | Overall | |
| Contiguous | | Non-Contiguous | |  | | | Contiguous | | | Non-Contiguous | | | | | |  |
| Contiguous | Non-Contiguous | | Contiguous | Non-Contiguous | |  | |  | | |  | |  | |  | | |  | | |  | | |  | | | | |  | | |
| **L** | 41/54  (75.93%) | 65/98  (66.33%) | | | 106/152  (69.74%) | 42/55  (76.36%) | | 51/77  (66.23% | | | | 93/132  (70.45%) | | 3247/4048  (80.21%) | | | 1678/2313  (72.55%) | | | | | 4925/6361  (77.42%) | | | 664/767  (86.57%) | | | 368/517  (71.18%) | | | | | | | 1032/1284  (80.37%) | |
| 36/66  (54.55%) | 29/32  (90.62%) | | 33/59  (55.93%) | 18/18  (100%) | |  | |  | | | 855/1309  (65.32%) | | 823/1004  (81.97%) | |  | | |  | | | 180/292  (61.64%) | | | | | 188/225  (83.56%) | |  | | | |
| **L I** | 43/54  (79.63%) | 62/98  (63.27%) | | | 105/152  (69.08%) | 44/55  (80%) | | 49/77  (63.64%) | | | | 93/132  (70.45%) | | 3352/4048  (82.81%) | | | 1641/2313  (70.95%) | | | | | 4993/6361  (78.49%) | | | 674/767  (87.87%) | | | 354/517  (68.47%) | | | | | | | 1028/1284  (80.06%) | |
| 33/66  (50%) | | 29/32  (90.62%) | 31/59  (52.54%) | | 18/18  (100%) | | 819/1309  (62.57%) | | 822/1004  (81.87%) | | 173/292  (59.25%) | | | | | 181/225  (80.44%) | |
| **L I D** | 44/54  (81.48%) | 64/98  (65.31%) | | | 108/152  (71.05%) | 45/55  (81.82%) | | 51/77  (66.23%) | | | | 96/132  (72.72%) | | 3368/4048  (83.20%) | | | 1633/2313  (70.60%) | | | | | 5001/6361  (78.62%) | | | 674/767  (87.87%) | | | 344/517  (66.54%) | | | | | | | 1018/1284  (79.28%) | |
| 35/66  (53.03%) | | 29/32  (90.62%) | 33/59  (55.93%) | 18/18  (100%) | |  | |  | | | 813/1309  (62.11%) | | 820/1004  (81.67%) | |  | | |  | | | 167/292  (57.19%) | | | | | 177/225  (78.67%) | |  | | | |
| **L I D R** | 39/54  (72.22%) | 78/98  (79.59%) | | | 117/152  (76.97%) | 40/55  (72.73%) | | 61/77  (79.22%) | | | | 101/132  (76.51%) | | 3307/4048  (81.69%) | | | 1784/2313  (77.13%) | | | | | 5091/6361  (80.03%) | | | 700/767  (91.26%) | | | 389/517  (75.24%) | | | | | | | 1089/12284  (84.81%) | |
| 48/66  (72.73%) | | 30/32  (93.75%) | 44/59  (74.58%) | 17/18  (94.44%) | |  | |  | | | 977/1309  (74.64%) | | 807/1004  (80.38%) | |  | | |  | | | 209/292  (71.58%) | | | | | 180/225  (80%) | |  | | | |
| **I** | 47/54  (87.04%) | 86/98  (87.76%) | | | 133/152  (87.50%) | 47/55  (85.45%) | | 70/77  (90.91%) | | | | 117/132  (88.63%) | | 3112/4048  (76.88%) | | | 1895/2313  (81.93%) | | | | | 5007/6361  (78.71%) | | | 575/767  (74.97%) | | | 415/517  (80.27%) | | | | | | | 990/1284  (77.10%) | |
| 62/66  (93.94%) | | 24/32  (75.00%) | 57/59  (96.61%) | 13/18  (72.22%) | |  | |  | | | 1078/1309  (82.35%) | | 817/1004  (81.37%) | |  | | |  | | | 234/292  (80.14%) | | | 181/225  (80.44%) | | | | |
| **I D** | 47/54  (87.04%) | 86/98  (87.76%) | | | 133/152  (87.50%) | 47/55  (85.45%) | | 70/77  (90.91%) | | | | 117/132  (88.63%) | | 3118/4048  (77.03%) | | | 1879/2313  (81.24%) | | | | | 4997/6361  (78.56%) | | | 579/767  (75.49%) | | | 414/517  (80.08%) | | | | | | | 993/1284  (77.34%) | |
| 62/66  (93.94%) | | 24/32  (75%) | 57/59  (96.61%) | 13/18  (90.91%) | |  | |  | | | 1067/1309  (81.51%) | | 812/1004  (80.88%) | |  | | |  | | | 234/292  (80.14%) | | | 180/225  (80%) | | | | |
| **I D R** | 44/54  (81.48%) | 77/98  (78.57%) | | | 121/152  (79.61%) | 45/55  (81.82%) | | 61/77  (79.22%) | | | | 106/132  (80.30%) | | 3304/4048  (81.62%) | | | 1739/2313  (75.18%) | | | | | 5043/6361  (79.28%) | | | 606/767  (79.01%) | | | 376/517  (72.73%) | | | | | | | 982/1284  (76.48%) | |
| 49/66  (74.24%) | | 28/32  (87.50%) | 44/59  (74.58%) | 17/18  (94.44%) | |  | |  | | | 967/1309  (73.87%) | | 772/1004  (76.89%) | |  | | |  | | | 201/292  (68.84%) | | | 175/225  (77.78%) | | | | |
| **D** | 41/54  (75.93%) | 65/98  (66.33%) | | | 106/152  (69.74%) | 42/55  (76.36%) | | 52/77  (67.53%) | | | | 94/132  (71.21%) | | 2969/4048  (73.34%) | | | 1236/2313  (53.44%) | | | | | 4205/6361  (66.11%) | | | 541/767  (70.53%) | | | 295/517  (57.06%) | | | | | | | 836/1284  (65.11%) | |
| 47/66  (71.21%) | | 18/32  (56.25%) | 42/59  (71.19%) | 10/18  (55.56%) | |  | |  | | | 799/1309  (61.04%) | | 437/1004  (43.53%) | |  | | |  | | | 182/292  (62.33%) | | | 113/225  (50.22%) | | | | |
| **D R** | 43/54  (79.63%) | 76/98  (77.55%) | | | 119/152  (78.29%) | 44/55  (80.00%) | | 60/77  (77.92%) | | | | 104/132  (78.78%) | | 3291/4048  (81.30%) | | | 1731/2313  (74.84%) | | | | | 5022/6361  (78.95%) | | | 609/767  (79.40%) | | | 374/517  (72.34%) | | | | | | | 983/1284  (76.56%) | |
| 47/66  (71.21%) | | 29/32  (90.62%) | 43/59  (72.88%) | 17/18  (94.44%) | |  | |  | | | 962/1309  (73.49%) | | 769/1004  (76.59%) | |  | | |  | | | 200/292  (68.49%) | | | 174/225  (77.33%) | | | | |
| **R** | 45/54  (83.33%) | 77/98  (78.57%) | | | 122/152  (80.26%) | 46/55  (83.64%) | | 61/77  (79.22%) | | | | 107/132  (81.06%) | | 3289/4048  (81.25%) | | | 1741/2313  (75.27%) | | | | | 5030/6361  (79.08%) | | | 609/767  (79.40%) | | | 374/517  (72.34%) | | | | | | | 983/1284  (76.56%) | |
| 48/66  (72.73%) | | 29/32  (90.62%) | 44/59  (74.58%) | 17/18  (94.44%) | |  | |  | | | 967/1309  (73.87%) | | 774/1004  (77.09%) | |  | | |  | | | 201/292  (68.84%) | | | 173/225  (76.89%) | | | | |

**Observations**

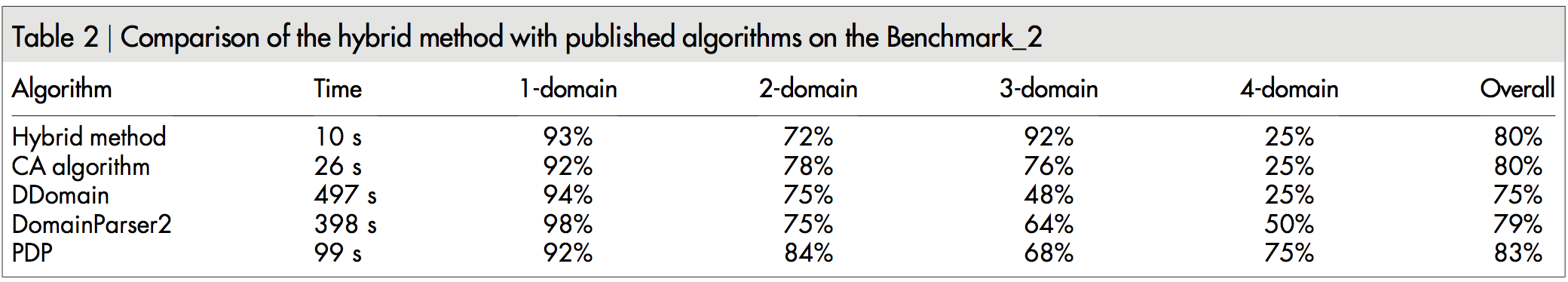
* We are currently employing all the L, I, D & R(marked in yellow) for single vs multi-domain classification. While this feature set works best for our self-created dataset but it doesn’t work quite well for Benchmark2, Benchmark3 and ASTRAL SCOP30 dataset.
* While it can be observed that for both I(green) & I D(pink) as feature set, the performance of the SVM has been far better on Benchmark2 & Benchmark3 and is very comparable for ASTRAL SCOP 30 as compared to when L, I, D & R were used as feature set.
* Only in the case of self-created dataset has the feature set of I & I D made the SVM perform poorly as compared to when L, I, D & R were used.

Comparison with other tools in identifying single domain proteins on different datasets:

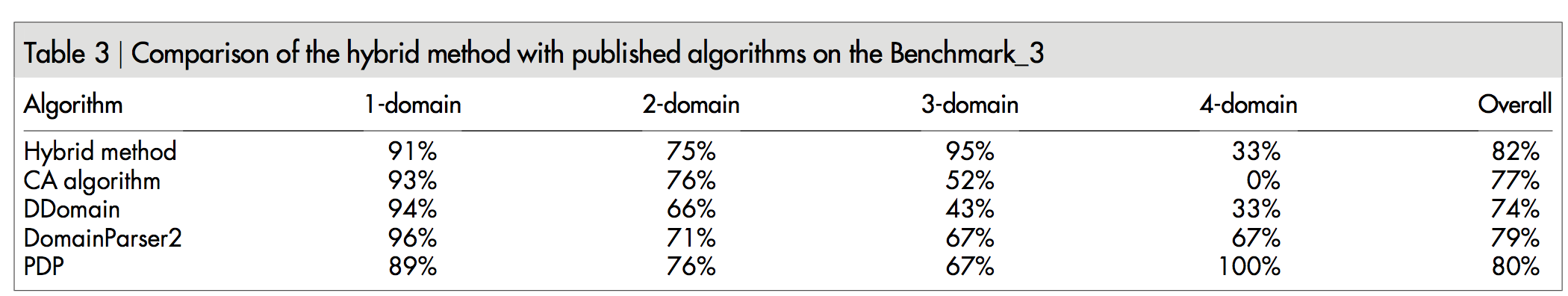
1. ASTRAL SCOP30

We correctly identified 77.03% of the 1-domain and 81.24% multi-domain proteins in ASTRAL SCOP30 dataset when I, D were used as feature set. Although, when L, I, D & R were used we correctly identified 81.69% of the single domain proteins and 77.13% of the multi-domain proteins.

2. Benchmark\_2

We correctly identified 87.04% of the 1-domain and 87.76% multi-domain proteins in Benchmark\_2 dataset when either of I or I, D were used as feature set. When L, I, D & R were used, we correctly identified 72.22% of the single domain proteins and 79.59% of the multi-domains.

3. Benchmark\_3

We correctly identified 85.45% of the 1-domain and 90.91% multi-domain proteins in Benchmark\_3 dataset when either of I or I, D were used as feature set. When L, I, D & R were used, we correctly identified 72.73% of the single domain proteins and 79.22% of the multi-domains.

**Conclusion:** Thus, we should employ a combination of Interaction Strength & Density as feature set when classifying single vs multi domain proteins.

**Update**: The formula we used for calculating Interaction Energy was:

[Nxy/(Nx+Ny)]x100

Where Nxy are the total number of interactions between the two clusters obtained by using k-means for k=2. Nx & Ny are the size of the two obtained clusters.

I tweaked the formula under the hypothesis that the number of inter-domain interactions are significantly lesser than the number of intra-domain contacts. While the formula mentioned above remains the same but the definition of Nx & Ny changes. Now, Nx & Ny represent the number of intra-domain interactions respectively in the two obtained clusters. Based on this, I updated the results where Interaction Energy was being used as a feature in the feature set.

For each column, I have highlighted the values as follows:

Feature set consisting of Interaction Energy have their updated values in Purple.

Highest percentage => Blue

Second highest percentage => Green

Third highest percentage => Yellow

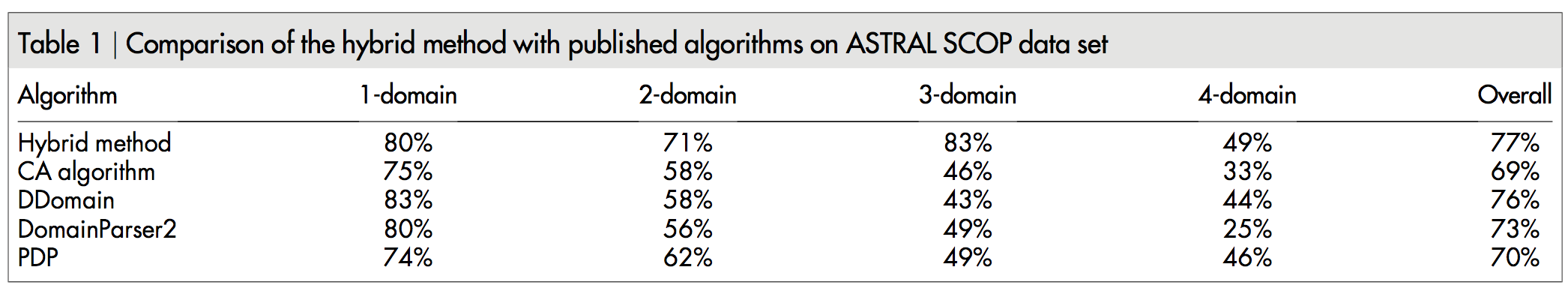
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature Set** | **Benchmark2** | | | | | | | **Benchmark3** | | | | | | | | | **ASTRAL SCOP30** | | | | | | | | | | **Self Created** | | | | | | | | | | | | | |
|  | Single Correct | Multi Correct | | | | Overall | Single Correct | | Multi Correct | | | | | Overall | | Single Correct | | | Multi Correct | | | | | | Overall | | | Single Correct | | | Multi Correct | | | | | | | | Overall | |
| Contiguous | | | Non-Contiguous | |  | | | Contiguous | | | Non-Contiguous | | | | | | |  |
| Contiguous | Non-Contiguous | | | Contiguous | Non-Contiguous | | |  | |  | | |  | | |  | |  | | |  | | |  | | |  | | | | | |  | | |
| **L** | 41/54  (75.93%) | 65/98  (66.33%) | | | | 106/152  (69.74%) | 42/55  (76.36%) | | 51/77  (66.23% | | | | 93/132  (70.45%) | | | 3247/4048  (80.21%) | | | 1678/2313  (72.55%) | | | | 4925/6361  (77.42%) | | | | | 664/767  (86.57%) | | | 368/517  (71.18%) | | | | | | 1032/1284  (80.37%) | | | |
| 36/66  (54.55%) | 29/32  (90.62%) | | | 33/59  (55.93%) | 18/18  (100%) | | | 855/1309  (65.32%) | | 823/1004  (81.97%) | | 180/292  (61.64%) | | | | 188/225  (83.56%) | |
| **L I** | 47/54  (87.04%) | 85/98  (86.73%) | | | | 132/152  (86.84%) | 48/55  (87.27%) | | 65/77  (84.42%) | | | | | 113/132  (85.61%) | | 3416/4048  (84.39%) | | | 1860/2313  (80.42%) | | | | | | 5276/6361  (82.94%) | | | 717/767  (93.48%) | | | 417/517  (80.66%) | | | | | | | | 1134/1284  (88.32%) | |
| 53/66  (80.30%) | | 32/32  (100.0%) | | 47/59  (79.66%) | | | 18/18  (100%) | | 1005/1309  (76.78%) | | | 855/1004  (85.16%) | | 224/292  (76.71%) | | | | | | 193/225  (85.78%) | |
| **L I D** | 47/54  (87.04%) | 84/98  (85.71%) | | | | 131/152  (86.18%) | 48/55  (87.27%) | | 65/77  (84.42%) | | | | 113/132  (85.61%) | | | 3477/4048  (85.89%) | | | 1848/2313  (79.90%) | | | | 5325/6361  (83.71%) | | | | | 715/767  (93.22%) | | | 411/517  (79.50%) | | | | | | 1126/1284  (87.69%) | | | |
| 52/66  (78.79%) | | 32/32  (100.0%) | | 47/59  (79.66%) | 18/18  (100%) | | | 994/1309  (75.94%) | | 854/1004  (85.06%) | | 218/292  (74.66%) | | | | 193/225  (85.78%) | |
| **L I D R** | 40/54  (74.07%) | 85/98  (86.73%) | | | | 125/152  (82.24%) | 41/55  (74.55%) | | 66/77  (85.71%) | | | | 107/132  (81.06%) | | | 3192/4048  (78.85%) | | | 1912/2313  (82.66%) | | | | 5104/6361  (80.24%) | | | | | 710/767  (92.57%) | | | 431/517  (83.37%) | | | | | | 1141/1284  (88.86%) | | | |
| 55/66  (83.33%) | | 30/32  (93.75%) | | 49/59  (83.05%) | 17/18  (94.44%) | | | 1067/1309  (81.51%) | | 845/1004  (84.16%) | | 238/292  (81.51%) | | | | 193/225  (85.78%) | |
| **I** | 44/54  (81.48%) | 90/98  (91.84%) | | | | 134/152  (88.16%) | 45/55  (81.82%) | | 73/77  (94.81%) | | | | 118/132  (89.39%) | | | 3087/4048  (76.26%) | | | 1978/2313  (85.52%) | | | | 5065/6361  (79.63%) | | | | | 579/767  (75.49%) | | | 429/517  (82.98%) | | | | | | | | 1008/1284  (78.50%) | |
| 62/66  (93.94%) | | 28/32  (87.50%) | | 58/59  (98.31%) | 15/18  (83.33%) | | | 1115/1309  (85.18%) | | 863/1004  (85.96%) | | 240/292  (82.19%) | 189/225  (84.0%) | | | | | |
| **I D** | 45/54  (83.33%) | 89/98  (90.82%) | | | | 134/152  (88.16%) | 46/55  (83.64%) | | 72/77  (93.51%) | | | | 118/132  (89.39%) | | | 3110/4048  (76.83%) | | | 1963/2313  (84.87%) | | | | 5073/6361  (79.75%) | | | | | 580/767  (75.62%) | | | 425/517  (82.21%) | | | | | | | | 1005/1284  (78.27%) | |
| 61/66  (92.42%) | | 28/32  (87.50%) | | 57/59  (96.61%) | 15/18  (83.33%) | | | 1108/1309  (84.64%) | | 855/1004  (85.16%) | | 236/292  (80.82%) | 189/225  (84.0%) | | | | | |
| **I D R** | 47/54  (87.04%) | 87/98  (88.78%) | | | | 134/152  (88.16%) | 47/55  (85.45%) | | 69/77  (89.61%) | | | | 116/132  (87.88%) | | | 3437/4048  (84.91%) | | | 1864/2313  (80.59%) | | | | 5301/6361  (83.34%) | | | | | 646/767  (84.22%) | | | 413/517  (79.88%) | | | | | | | | 1059/1284  (82.48%) | |
| 57/66  (86.36%) | | 30/32  (93.75%) | | 53/59  (89.83%) | 16/18  (88.89%) | | | 1035/1309  (79.07%) | | 829/1004  (82.57%) | | 225/292  (77.05%) | 188/225  (83.56%) | | | | | |
| **D** | 41/54  (75.93%) | 65/98  (66.33%) | | | | 106/152  (69.74%) | 42/55  (76.36%) | | 52/77  (67.53%) | | | | 94/132  (71.21%) | | | 2969/4048  (73.34%) | | | 1236/2313  (53.44%) | | | | 4205/6361  (66.11%) | | | | | 541/767  (70.53%) | | | 295/517  (57.06%) | | | | | | | | 836/1284  (65.11%) | |
| 47/66  (71.21%) | | 18/32  (56.25%) | | 42/59  (71.19%) | 10/18  (55.56%) | | | 799/1309  (61.04%) | | 437/1004  (43.53%) | | 182/292  (62.33%) | 113/225  (50.22%) | | | | | |
| **D R** | 43/54  (79.63%) | 76/98  (77.55%) | | | | 119/152  (78.29%) | 44/55  (80.00%) | | 60/77  (77.92%) | | | | 104/132  (78.78%) | | | 3291/4048  (81.30%) | | | 1731/2313  (74.84%) | | | | 5022/6361  (78.95%) | | | | | 609/767  (79.40%) | | | 374/517  (72.34%) | | | | | | | | 983/1284  (76.56%) | |
| 47/66  (71.21%) | | 29/32  (90.62%) | | 43/59  (72.88%) | 17/18  (94.44%) | | | 962/1309  (73.49%) | | 769/1004  (76.59%) | | 200/292  (68.49%) | 174/225  (77.33%) | | | | | |
| **R** | 45/54  (83.33%) | 77/98  (78.57%) | | | | 122/152  (80.26%) | 46/55  (83.64%) | | 61/77  (79.22%) | | | | | 107/132  (81.06%) | | 3289/4048  (81.25%) | | | 1741/2313  (75.27%) | | | | | | 5030/6361  (79.08%) | | | 609/767  (79.40%) | | | 374/517  (72.34%) | | | | | | | | 983/1284  (76.56%) | |
| 48/66  (72.73%) | | | 29/32  (90.62%) | 44/59  (74.58%) | | 17/18  (94.44%) | | | 967/1309  (73.87%) | 774/1004  (77.09%) | | | | | 201/292  (68.84%) | | | 173/225  (76.89%) | | | | |

**Observations:**

* The updated value of Interaction Strength makes the combination LI/LID as the best feature set to identify single domains from multi. Please note that both of these feature sets perform almost equally well.
* Along with correctly identifying Single domain proteins, the LI/LID combination is quite successful in identifying multi-domain non-contiguous chains as well. It has performed at 100% for Benchmark2 & Benchmark3 and >85% for ASTRAL SCOP & Self-created dataset in identifying non-contiguous multi-domain chains.
* While the LI/LID set performs best in identifying single and non-contiguous multi-domain proteins, multi-domain identification is best performed by I/ID feature set.
* Thus, the set I/ID is best when it comes to identifying multi-domain proteins and the set LI/LID is best when it comes to identifying single domain proteins.
* Much heed should not be paid to the overall accuracy, as it depends on the number of single and multi-domain chains present in the given test dataset. For example, a dataset consisting of too many single-domain proteins will have a higher overall accuracy if the feature set LI/LID is used, as is the case with the ASTRAL SCOP30 dataset. While, if the number of multi-domain proteins are significantly higher in a dataset then the feature set I/ID will perform better overall as in Benchmark2 & Benchmark3.

Comparison with other tools in identifying single domain proteins on different datasets. Since single domain identification is the key performance indicator, I’m comparing the values when LID feature set is used. When the previous formula of calculating Interaction Strength was used, we used ID as our feature set as it was giving the best results. I’m giving a comparison of our previous best results as well.

1. ASTRAL SCOP30

Previous formula used to evaluate I & using ID as feature set:

1-domain => 77.03%

multi-domain => 81.24%

Overall => 78.56%

New formula used to evaluate I & using LID as feature set:

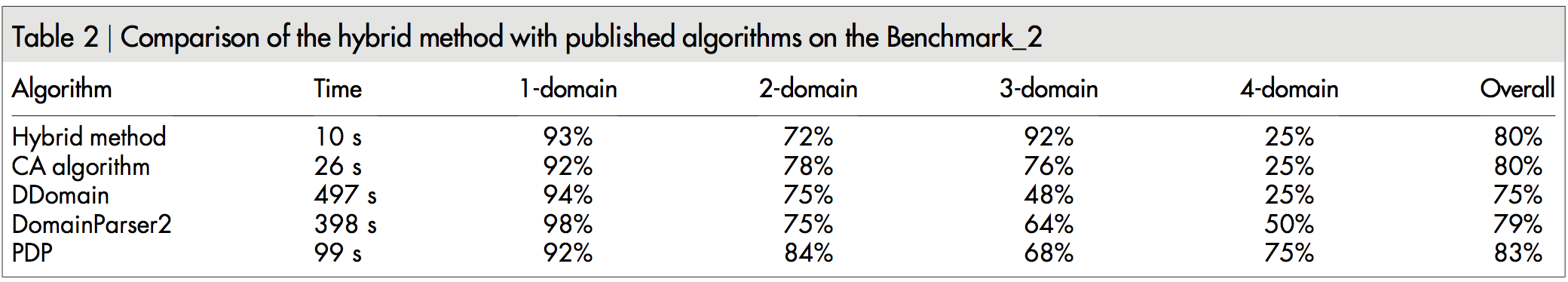
1-domain => 85.89%

multi-domain => 79.90%

Overall => 83.71%

**Observation:** We have topped all of the above methods here to identify single domain proteins by outscoring the best program DDomain which correctly identified 83% of the single domain proteins.

2. Benchmark\_2



Previous formula used to evaluate I & using ID as feature set:

1-domain => 47/54: 87.04%

multi-domain => 86/98: 87.76%

Overall => 133/152: 87.50%

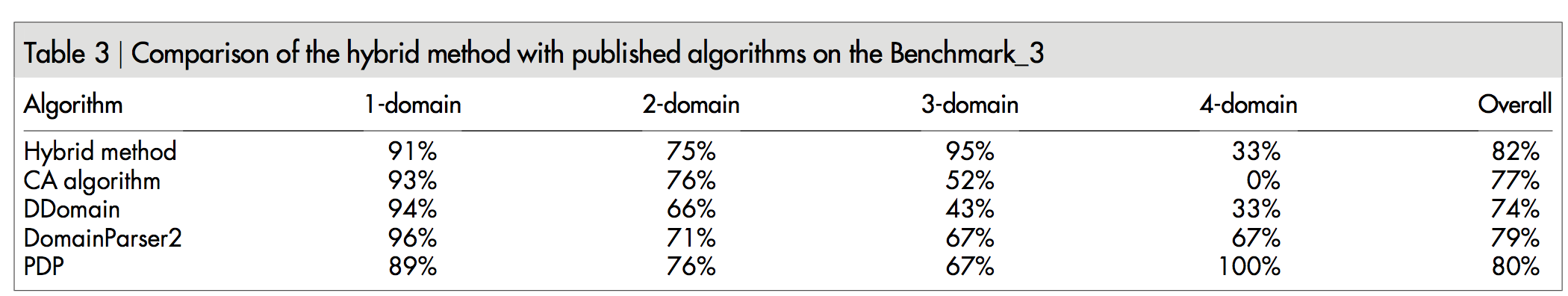
New formula used to evaluate I & using LID as feature set:

1-domain => 47/54: 87.04%

multi-domain => 84/98: 85.71%

Overall => 131/152: 86.18%

3. Benchmark\_3



Previous formula used to evaluate I & using ID as feature set:

1-domain => 47/55: 85.45%

multi-domain => 70/77: 90.91%

Overall => 117/132: 88.63%

New formula used to evaluate I & using LID as feature set:

1-domain => 48/55: 87.27%

multi-domain => 65/77: 84.42%

Overall => 113/132: 85.61%

**Observation**: For Benchmark\_2 & Benchmark\_3, the performances are almost similar. But, there are improvements when we use ID as feature set with the newly calculated value of I. Please refer to the table for the same.

4. Self-created dataset

Previous formula used to evaluate I & using ID as feature set:

1-domain => 579/767: 75.49%

multi-domain => 414/517: 80.08%

Overall => 993/1284: 77.34%

New formula used to evaluate I & using LID as feature set:

1-domain => 715/767: 93.22%

multi-domain => 411/517: 79.50%

Overall => 1126/1284: 87.69%

**Observation**: There is a huge improvement in identification of single domain proteins and the multi-domain identification is also comparable.

Based on the above set of observations, we can conclude that the updated formula for calculating the Interaction Strength gives better results and it is more intuitive as it supports the hypothesis that the number of inter-domain interactions should be much lesser than intra-domain interactions. Along with this, for the first phase SVM, where we are trying to classify the single-domain & multi-domain proteins the best feature set to use is LID.