1 Supplementary materials

1.1 NR and Clustering Descriptive statistics

Summary statistics of protein clusters, unclustered proteins, proteins, amino acid content and taxonomic assignment for CD-Hit clustering at 95% and then clustering of representative sequences a 90%, 85%, 80%, 75%, 70% and 65% identity, respectively is shown in Table 1.

Table 1: The nr95, nr90, ..., nr65 summary statistics.

Similarity	Clusters	Unclustered proteins	Proteins	#Amino Acids	#taxa
95	88,276,733	64,018,330	174,233,775	63,580,247,140	744,237
90	72,393,597	63,135,775	88,276,733	30,462,708,282	231,410
85	61,757,844	54,844,642	72,393,597	24,951,507,349	155,471
80	53,308,820	47,596,227	61,757,844	21,255,665,717	123,299
75	46,157,068	41,246,925	53,308,820	18,319,406,051	106,039
70	40,010,418	35,764,294	46,157,068	15,844,729,956	95,704
65	34,451,690	30,647,959	40,010,418	13,712,952,017	88,850

1.2 BoaG queries

Following are the BoaG queries that used in this work to retrieve annotations from protein sequences and clusters. The output of these two queries in Fig. 1 and Fig. 2 were fed into algorithm 1 and algorithm 2 in the paper.

```
1  s: Sequence = input;
2  count : output sum [string][string] of int;
3  foreach(i:int; def(s.annotation[i]))
4    count [s.seqid][s.annotation[i].tax_id] << 1;
```

```
#Following are few lines of the output
count[12E8][9913] = 6
count[1BZW][3818] = 44
count[1C08][83333] = 4
```

Figure 1: Frequency of taxonomic assignment for identifying taxonomically misclassified protein sequences in the NR database. In line 1, s is defined a Sequence type. Line 2, count is an output aggregator that produces the sum of the output indexed over protein sequence ID and taxonomic ID.

```
s: Sequence = input;
clstrOut : output sum [int][string][string] of int;
foreach(i:int; def(s.annotation[i]))
  foreach(j:int; def(s.cluster[j]))
    clstrOut[s.cluster[j].similarity][s.cluster[j].cid] [s.annotation
    [i].tax_name]<< 1;
```

```
#Following are few lines of output
clstrOut[95][9667021][Mycoplasma gallisepticum] = 20
clstrOut[95][9667086][Lactobacillus helveticus] = 60
clstrOut[95][10025876][Okarito brown kiwi] = 21
```

Figure 2: Frequencies of taxonomic assignments for each cluster in different similarity level. The variable clstrOut is output aggregators that produces the sum of output indexed over similarity level, cluster id, and taxonomic name

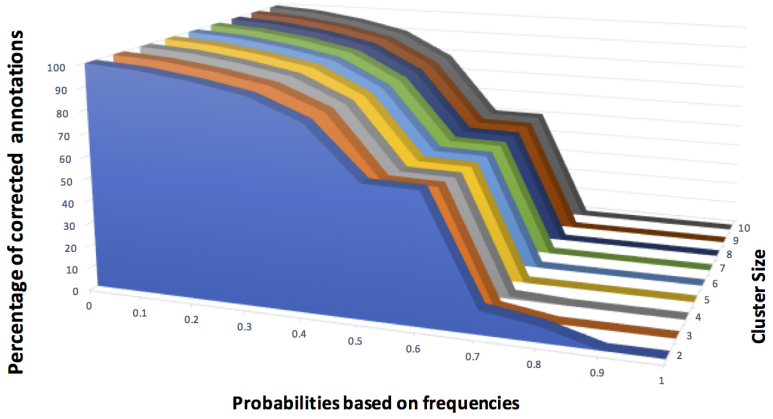

Figure 3: The input parameters to the sensitivity analysis are probability based on the annotation frequencies and the cluster size. The Z-axis shows for what percentages of the misclassified sequences, our approach can propose taxonomic assignment.

1.3 Sensitivity analysis

1.4 Misclassified sequences at clusters

Table 2 shows the number of potential misclassified sequences in the clusters at 95% similarity at the superkingdom level to the family level that has been detected by applying distance in the phylogenetic tree. The second column shows the number of entire clusters that have a certain number of taxonomic assignments.

Table 2: Misclassification in clusters of the NR database at 95% similarity.

# taxa	total	Root	kingdom	Phylum	Class	Order	Family
2	12,960,476	17,099	92,526	263,844	100,560	267,251	461,795
3	4,683,663	9,825	39,940	153,678	63,414	153,996	291,418
4	2,328,246	7,314	25,361	95,038	33,603	102,671	173,810
5	1,293,767	5,136	17,915	56,510	22,253	62,025	101,675
6	566,574	4,936	14,660	39,410	15,738	46,913	66,741
7	566,574	3,652	13,642	23,206	12,160	40,760	49,046
8	403,513	2,719	8,433	10,622	10,577	24,259	36,463
9	289,289	1,635	6,655	7,291	8,890	19,608	28,549
10	235,451	1,423	4,744	8,991	8,586	16,070	22,026
¿10	1,832,313	22,921	63,513	3,951	65,196	155,804	200,642

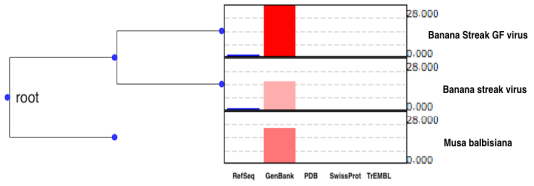

Figure 4: Misclassification detected in the cluster ID 490503 at 95% similarity of the NR dataset.

Figure 4 shows an example of detected taxonomically assigned annotations in the cluster-id 490503. Leaves are annotated with databases of origins and frequency of taxonomic assignments as a bar chart from all reviewed and unreviewed databases, i.e., RefSeq, GenBank, PDB UniProt\SwissProt, and UniProt\TrEMBL respectively. For this cluster, there is no annotation from PDB, SwissProt, and TrEMBL databases.

1.5 Correcting Taxonomic Misclassification

Figure 5 shows different percentages of conflicts from the subset of one million sequences in the NR database.

1.6 Case study: Glycine

In addition to the simulated and real-world datasets, we also explored more deeply some clusters that had *Glycine* to identify a very small subset of clusters that had a high probability of containing sequences with a taxonomic misclassification. We did this by making some basic assumptions and applying very stringent filters to pull out five clusters and explore them in detail. The five clusters we identified were at the 95% similarity level and had at least one protein with the genus *Glycine* in the cluster, and at least one other genus but not more than eight different genera and the proteins were not functionally annotated as *ribosomal proteins*. The cluster ids at 95% similarity

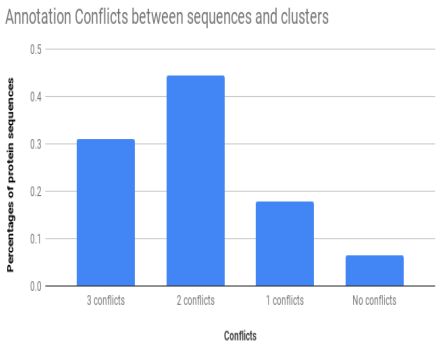

Figure 5: Conflicts between the set of top three taxonomic assignment in sequences and the top three taxonomic assignment in the clusters of the NR database.

that were identified are 13741561, 26559833, 37311095, 64827049 and 81606138. Each one of these clusters happened to contain a single protein sequence that was taxonomically classified in the genus *Glycine*.

These protein sequence ids are AAL67577, KHN16331, KHN42910, KHN24658, and KHN25027. The remaining protein sequence ids in these clusters were taxonomically classified with bacterial genera or synthetic construct. These protein sequence ids grouped by clusterid (protein ids) are:

- 13741561 (2GGD, 2PQB, 2PQD, AAB36219, AAL67577, ABY76172, AEP17820, AII71485, AMM72836, ARW80139, ARW80140, CBG92008, KIF05476, Q9R4E4, WP_037405162, WP_050745479, WP_117368106)
- 26559833 (EEV24075, KHN25027, WP_062335070, WP_076776357, WP_101965181)
- 37311095 (KHN24648, WP_007116868, WP_062330154, WP_065252519, WP_100269819)
- 64827049 (KHN42910, PZO92494, WP_007115888, WP_060994472, WP_062331401, WP_065253003, WP_065264285, WP_101963455)
- and 81606138 (KHN16331, WP_007115622, WP_060994515, WP_065262776)

In cluster 13741561, the one protein with a *Glycine* taxonomic assignment is AAL67577 and has a length of 455 AA. It aligns with the cluster's representative sequence from amino acid 77 to 455. The protein sequence AEP17820 is also in the cluster and aligns to the representative sequence between 1 and 76, and it is annotated as an expression vector. Many other sequences in the cluster are labeled as synthetic constructs suggesting AAL67577 may be a synthetic construct created by recombinant DNA techniques. Exploring AAL67577 further using NCBI's new SMARTBLAST tool reveals that it is most similar to proteins in bacterial species with almost full coverage of the alignment length. A strong indication that this protein is indeed misclassified. More details about the clusters, sequences, and the SMARTBLAST output are shown in our repository.