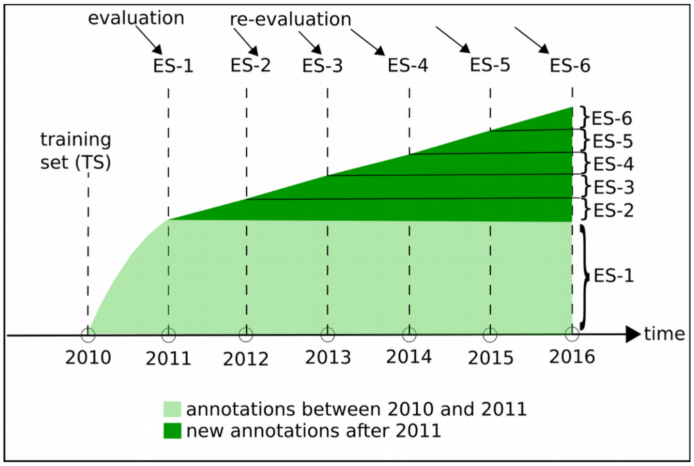

	MFO		BPO		CCO	
TS/ES	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
TS	22,715	3,252	23,124	7,549	24,216	1,224
ES-1	3,546	4,639	3,669	7,439	3,160	4,146
ES-2	346	448	540	1,088	235	297
ES-3	202	290	445	1,553	191	333
ES-4	136	184	420	1,015	174	241
ES-5	136	294	411	2,052	172	548
ES-6	133	199	398	1,067	162	298

Species: 9606, 10090, 3702, 10116, 559292, 9913, 284812, 83333, 224308, 44689

# of proteins and GO terms in training sets (UniProt/SwissProt time point 2010_01)						
	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	5,252	1,450	3,567	2,700	4,446	552
10090	3,682	1,104	4,265	3,588	3,847	446
3702	1,876	685	2,374	1,126	2,848	185
10116	2,253	1,252	2,327	1,955	1,962	374
559292	0	0	0	0	0	0
9913	103	80	62	85	65	38
284812	0	0	0	0	0	0
83333	786	12	12	10	160	6
224308	0	0	0	0	0	0
44689	284	184	469	353	474	112

# of proteins and GO terms in evaluation set 1 (ES-1)
(UniProt/SwissProt time points 2010_01 and 2011_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	16	24	8	16	16	25
10090	4	3	2	3	1	1
3702	10	11	21	36	22	37
10116	0	0	0	0	1	5
559292	0	0	0	0	0	0
9913	3	1	0	0	0	0
284812	0	0	0	0	0	0
83333	1	1	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	1	1

# of proteins and GO terms in evaluation set 2 (ES-2)
(based on ES-1 and UniProt/SwissProt time point 2012_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	2	4	3	3	5	8
10090	0	0	1	8	0	0
3702	1	1	3	4	2	1
10116	0	0	0	0	0	0
559292	0	0	0	0	0	0
9913	0	0	0	0	0	0
284812	0	0	0	0	0	0
83333	0	0	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	1	1

# of proteins and GO terms in evaluation set 3 (ES-3)
(based on ES-2 and UniProt/SwissProt time point 2013_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	1	1	3	7	4	4
10090	0	0	1	1	0	0
3702	0	0	3	5	2	14
10116	0	0	0	0	0	0
559292	0	0	0	0	0	0
9913	0	0	0	0	0	0
284812	0	0	0	0	0	0
83333	0	0	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	0	0

# of proteins and GO terms in evaluation set 4 (ES-4)
(based on ES-3 and UniProt/SwissProt time point 2014_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	1	3	3	5	4	8
10090	0	0	1	8	0	0
3702	0	0	3	5	2	1
10116	0	0	0	0	0	0
559292	0	0	0	0	0	0
9913	0	0	0	0	0	0
284812	0	0	0	0	0	0
83333	0	0	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	0	0

# of proteins and GO terms in evaluation set 5 (ES-5)
(based on ES-4 and UniProt/SwissProt time point 2015_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	1	2	3	10	4	12
10090	0	0	1	2	0	0
3702	0	0	3	9	2	15
10116	0	0	0	0	0	0
559292	0	0	0	0	0	0
9913	0	0	0	0	0	0
284812	0	0	0	0	0	0
83333	0	0	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	0	0

# of proteins and GO terms in evaluation set 6 (ES-6)
(based on ES-5 and UniProt/SwissProt time point 2016_01)

	MFO		BPO		CCO	
Organism	Proteins	GO terms	Proteins	GO terms	Proteins	GO terms
9606	1	3	3	7	4	10
10090	0	0	1	8	0	0
3702	0	0	3	4	2	1
10116	0	0	0	0	0	0
559292	0	0	0	0	0	0
9913	0	0	0	0	0	0
284812	0	0	0	0	0	0
83333	0	0	0	0	0	0
224308	0	0	0	0	0	0
44689	0	0	0	0	0	0

Training dataset, TS
Evaluation dataset, ES-1
Reevaluation datasets, ES-2, ES-3, …,ES-6

- ES-2, ES-3, …, and, ES-6 are created based on the growth of UniProtKB/SwissProt database.

Reevaluation datasets, ES-2', ES-3', …, ES-6'

- ES'-2, ES'-3, …, and , ES'-6 are created based on the monthly growth of TrEMBL database.
- This enables us to collect larger number of proteins for each set.

Creating Training Dataset, TS

Command: `python xTract_trainingSet -I1=uniprot_sprot.dat.2010_01`
Input file: UniProtKB/SwissProt file January 2010
Output files:

Three pairs of output files – one FASTA and one corresponding map file, a pair for each of MFO, BPO, and CCO ontological categories. Each FASTA file contains the protein sequences in FASTA format. On the other hand, the map file contains the sequence id used in the FASTA file and GO annotations that define the functions of the protein.

(1) MFO
```
1. uniprot_sprot.dat.2010_01.tfa_mfo.1
2. uniprot_sprot.dat.2010_01.tfa_mfo.1.map
```
(2) BPO
```
1. uniprot_sprot.dat.2010_01.tfa_bpo.1
2. uniprot_sprot.dat.2010_01.tfa_bpo.1.map
```
(3) CCO
```
1. uniprot_sprot.dat.2010_01.tfa_cco.1
2. uniprot_sprot.dat.2010_01.tfa_cco.1.map
```

Creating Evaluation Datasets: ES-1, ES-2, ES-3, ES-4, ES-5, and ES-6

ES-1. This dataset is used for evaluation of the prediction models. The set has those sequences that did not have Exp validation in January 2010 but gained such validation in January 2011.

Command:
```
python xTract_testSet -I1=uniprot_sprot.dat.2010_01 -I2=uniprot_sprot.dat.2011_01
```
Input files:

(1) UniProtKB/SwissProt file January 2010: `uniprot_sprot.dat.2010_01`
(2) UniProtKB/SwissProt file January 2011: `uniprot_sprot.dat.2011_01`

Output files:

Three pairs of output files – one FASTA and one corresponding map file, a pair for each of MFO, BPO, and CCO ontological categories. Each FASTA file contains the protein sequences in FASTA format. On the other hand, the map file contains the sequence id used in the FASTA file, the corresponding protein name, and GO annotations that define the functions of the protein.

(1) MFO
1. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_bpo.1`
2. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_bpo.1.map`

(2) BPO
1. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_mfo.1`
2. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_mfo.1.map`

(3) CCO
1. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_cco.1`
2. `uniprot_sprot.dat.2010_01-2011_01.tfa_LK_cco.1.map`