PubMed Text Similarity Model and its application to curation efforts in the Conserved Domain Database

Rezarta Islamaj, W. John Wilbur, Natalie Xie, Noreen R. Gonzales, Narmada Thanki, Roxanne Yamashita, Chanjuan Zheng, Aron Marchler-Bauer, and Zhiyong Lu

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894

**Supplementary file: Results-of-reference-suggestion**

 Table S1 Articles suggested and accepted as references for the CDD records in our dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CDD-id | Number of referenced articles in the CDD Record | References suggested by text mining | Number of articles accepted as a reference | Number of sentences w/ suggested references | Number of sentences with an accepted reference  |
| 100007 | 14 | 6 | 6 | 4 | 4 |
| 107251 | 9 | 6 | 5 | 3 | 3 |
| 132757 | 10 | 6 | 6 | 4 | 4 |
| 133064 | 7 | 4 | 3 | 4 | 4 |
| 143204 | 7 | 7 | 2 | 5 | 4 |
| 153288 | 17 | 14 | 13 | 10 | 10 |
| 153373 | 9 | 9 | 4 | 8 | 6 |
| 185765 | 12 | 9 | 8 | 12 | 12 |
| 187723 | 8 | 7 | 0 | 3 | 0 |
| 188681 | 6 | 6 | 6 | 5 | 4 |
| 188701 | 9 | 9 | 8 | 8 | 5 |
| 193545 | 7 | 5 | 3 | 8 | 6 |
| 198264 | 6 | 5 | 5 | 12 | 12 |
| 198339 | 12 | 11 | 10 | 7 | 7 |
| 200511 | 13 | 11 | 9 | 9 | 9 |
| 200538 | 11 | 5 | 5 | 8 | 8 |
| 206675 | 9 | 9 | 7 | 12 | 10 |
| 238367 | 4 | 3 | 2 | 4 | 2 |
| 238633 | 1 | 1 | 1 | 3 | 3 |
| 239094 | 4 | 4 | 2 | 5 | 4 |
| 239709 | 18 | 12 | 11 | 5 | 5 |
| 239798 | 5 | 5 | 3 | 2 | 2 |
| 239829 | 8 | 8 | 5 | 7 | 7 |
| 239981 | 9 | 7 | 4 | 8 | 5 |
| 240878 | 5 | 5 | 3 | 3 | 2 |
| 259871 | 7 | 6 | 3 | 6 | 5 |
| 269858 | 23 | 17 | 16 | 7 | 7 |
| 270060 | 13 | 8 | 7 | 17 | 14 |
| 270530 | 5 | 5 | 3 | 5 | 5 |
| 270805 | 7 | 7 | 6 | 9 | 7 |
| 276854 | 16 | 12 | 8 | 13 | 10 |
| 277120 | 7 | 6 | 5 | 9 | 6 |
| 277224 | 13 | 13 | 6 | 8 | 8 |
| 320362 | 36 | 28 | 26 | 10 | 9 |
| 320369 | 20 | 15 | 13 | 8 | 8 |
| 99806 | 6 | 6 | 3 | 4 | 3 |
| 99952 | 22 | 15 | 13 | 5 | 5 |
| TOTAL | **395** | **312** | **240** | **260** | **225** |
|  |  |  |  |  |  |
| ACCURACY |  | **78.99** | **76.92** |  | **86.54** |

 Table S2 The Discovery articles and their inclusion as new references.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CDD-id | Number of full text articles suggested | Number of more recent articles | Number of articles accepted as a reference | Number of recent articles accepted as reference |
| 100007 | 32 | 23 | 12 | 8 |
| 107251 | 29 | 28 | 8 | 8 |
| 132757 | 35 | 23 | 2 | 0 |
| 133064 | 35 | 33 | 7 | 6 |
| 143204 | 46 | 41 | 22 | 20 |
| 153288 | 73 | 56 | 43 | 30 |
| 153373 | 59 | 44 | 9 | 9 |
| 185765 | 64 | 41 | 45 | 33 |
| 187723 | 29 | 24 | 2 | 0 |
| 188681 | 48 | 33 | 1 | 1 |
| 188701 | 60 | 45 | 41 | 34 |
| 193545 | 69 | 63 | 17 | 16 |
| 198264 | 107 | 63 | 81 | 48 |
| 198339 | 54 | 38 | 35 | 27 |
| 200511 | 73 | 56 | 32 | 22 |
| 200538 | 48 | 28 | 12 | 6 |
| 206675 | 90 | 68 | 56 | 40 |
| 238367 | 36 | 34 | 9 | 8 |
| 238633 | 27 | 23 | 2 | 2 |
| 239094 | 48 | 42 | 9 | 8 |
| 239709 | 43 | 37 | 33 | 29 |
| 239798 | 18 | 14 | 12 | 9 |
| 239829 | 48 | 33 | 36 | 23 |
| 239981 | 69 | 62 | 42 | 41 |
| 240878 | 26 | 14 | 19 | 9 |
| 259871 | 50 | 32 | 29 | 24 |
| 269858 | 63 | 25 | 50 | 18 |
| 270060 | 117 | 43 | 59 | 21 |
| 270530 | 27 | 17 | 10 | 8 |
| 270805 | 86 | 73 | 28 | 27 |
| 276854 | 112 | 32 | 66 | 20 |
| 277120 | 66 | 43 | 12 | 8 |
| 277224 | 75 | 28 | 34 | 12 |
| 320362 | 89 | 34 | 57 | 23 |
| 320369 | 69 | 27 | 63 | 23 |
| 99806 | 39 | 31 | 23 | 18 |
| 99952 | 41 | 27 | 33 | 23 |
| **TOTAL** | **2100** | **1378** | **1051** | **662** |
|  |  |  |  |  |
| **Accuracy[[1]](#footnote-1)** |  |  | **0.50** | **0.48** |

1. We anticipate that the accuracy will improve with further optimizations, as discussed in the paper. [↑](#footnote-ref-1)