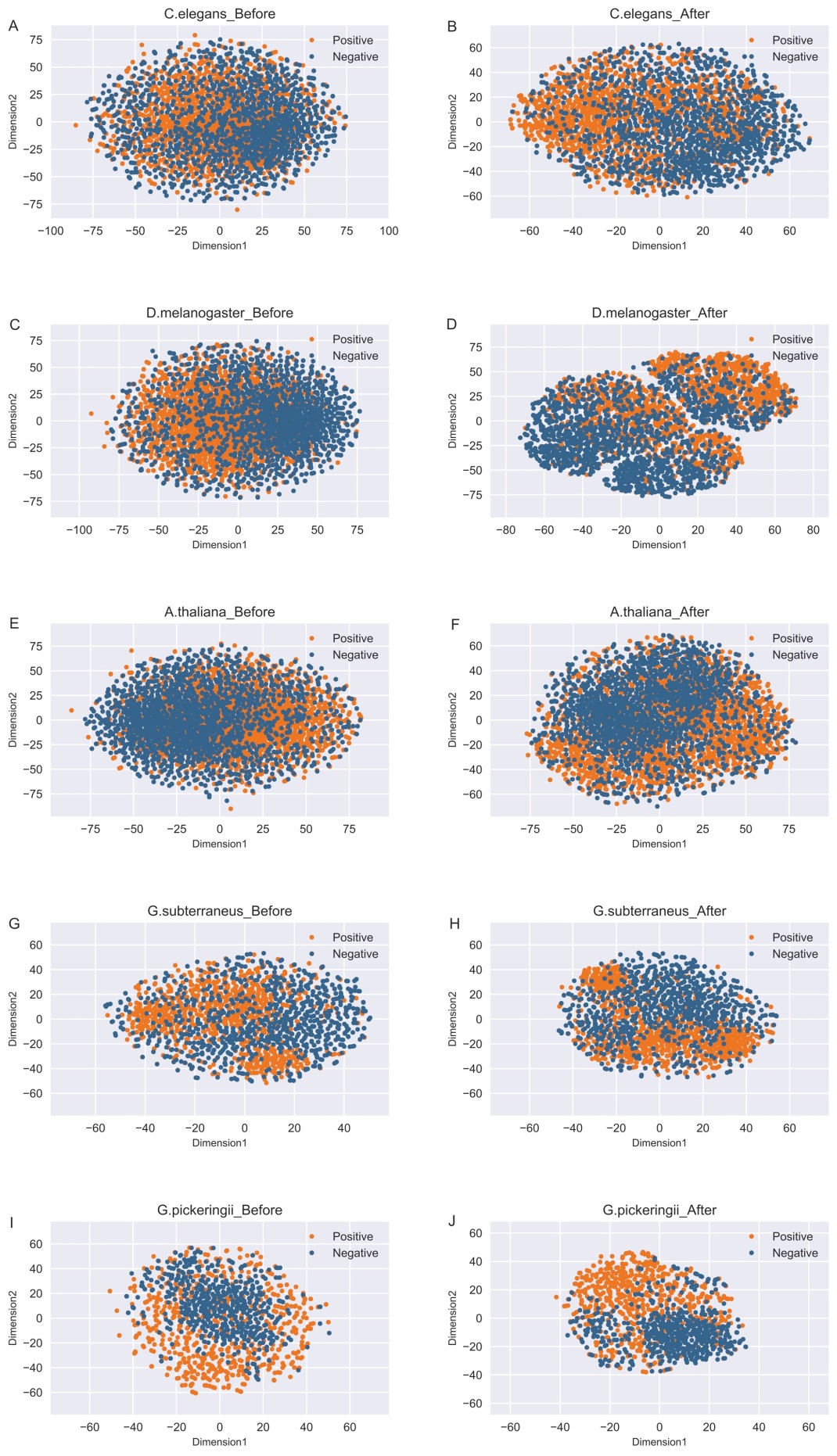
**Supporting Information**

**Supplementary methods**

**Procedure of 10-fold validation method**

The procedure of 10-fold cross validation method is described in the following steps. First, a dataset is randomly partitioned into 10 subsets with an equal size. Of the 10subsets, 9 subsets are chosen as the training data to train a predictive model, while the remaining single subset is retained as the validation data to test the model. This process is then repeated 10times, with each of the 10 subsets used exactly once as the validation data. Lastly, the ten results are averaged to obtain a final prediction estimation.

**Supplementary figures**

****

**Figure S1. T-SNE visualization of the five datasets in feature space before feature optimization and after feature optimization.**

**Supplementary tables**

**Table S1.** Performance of the three kernels of SVM.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Kernels** | **Datasets** | **Acc** | **SN** | **SP** | **MCC** | **TP** | **FN** | **FP** | **TN** |
| RBF | *C.elegans* | 0.786 | 0.791 | 0.782 | 0.573 | 1229 | 325 | 339 | 1215 |
| *D.melanogaster* | 0.815 | 0.841 | 0.788 | 0.630 | 1488 | 281 | 375 | 1394 |
| *A.thaliana* | 0.764 | 0.768 | 0.761 | 0.529 | 1519 | 459 | 473 | 1505 |
| *E.coli* | 0.787 | 0.778 | 0.796 | 0.575 | 302 | 86 | 79 | 309 |
| *G.subterruneus* | 0.817 | 0.830 | 0.804 | 0.634 | 751 | 154 | 177 | 728 |
| *G.pickeringii* | 0.832 | 0.838 | 0.826 | 0.664 | 477 | 92 | 99 | 470 |
| Linear | *C.elegans* | 0.759 | 0.773 | 0.745 | 0.518 | 1201 | 353 | 396 | 1158 |
| *D.melanogaster* | 0.792 | 0.804 | 0.780 | 0.584 | 1423 | 346 | 389 | 1380 |
| *A.thaliana* | 0.754 | 0.750 | 0.757 | 0.507 | 1484 | 494 | 481 | 1497 |
| *E.coli* | 0.754 | 0.755 | 0.753 | 0.508 | 293 | 95 | 96 | 292 |
| *G.subterruneus* | 0.785 | 0.793 | 0.776 | 0.569 | 718 | 187 | 203 | 702 |
| *G.pickeringii* | 0.786 | 0.793 | 0.780 | 0.573 | 451 | 118 | 125 | 444 |
| Poly | *C.elegans* | 0.723 | 0.818 | 0.627 | 0.454 | 1271 | 283 | 579 | 975 |
| *D.melanogaster* | 0.581 | 0.178 | 0.983 | 0.272 | 315 | 1454 | 30 | 1739 |
| *A.thaliana* | 0.701 | 0.846 | 0.556 | 0.420 | 1673 | 305 | 878 | 1100 |
| *E.coli* | 0.697 | 0.933 | 0.461 | 0.447 | 362 | 26 | 209 | 179 |
| *G.subterruneus* | 0.670 | 0.957 | 0.383 | 0.415 | 866 | 39 | 558 | 347 |
| *G.pickeringii* | 0.717 | 0.949 | 0.485 | 0.490 | 540 | 29 | 293 | 276 |
| Sigmod | *C.elegans* | 0.755 | 0.730 | 0.781 | 0.512 | 1135 | 419 | 341 | 1213 |
| *D.melanogaster* | 0.800 | 0.821 | 0.778 | 0.600 | 1453 | 316 | 392 | 1377 |
| *A.thaliana* | 0.758 | 0.766 | 0.750 | 0.516 | 1515 | 463 | 495 | 1483 |
| *E.coli* | 0.768 | 0.786 | 0.75 | 0.536 | 305 | 83 | 97 | 291 |
| *G.subterruneus* | 0.797 | 0.811 | 0.783 | 0.595 | 734 | 171 | 196 | 709 |
| *G.pickeringii* | 0.825 | 0.794 | 0.856 | 0.651 | 452 | 117 | 82 | 487 |

**Table S2.** Performances of the original features (Before optimization) and the optimal features (After optimization).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | | **Acc** | **SN** | **SP** | **MCC** | **TP** | **FN** | **FP** | **TN** |
| *C.elegans* | Before | 0.786 | 0.791 | 0.782 | 0.573 | 1229 | 325 | 339 | 1215 |
| After | 0.815 | 0.824 | 0.807 | 0.631 | 1280 | 274 | 300 | 1254 |
| *D.melanogaster* | Before | 0.815 | 0.841 | 0.788 | 0.630 | 1488 | 281 | 375 | 1394 |
| After | 0.830 | 0.838 | 0.822 | 0.661 | 1483 | 286 | 314 | 1455 |
| *A.thaliana* | Before | 0.764 | 0.768 | 0.761 | 0.529 | 1519 | 459 | 473 | 1505 |
| After | 0.787 | 0.778 | 0.796 | 0.573 | 1538 | 440 | 404 | 1574 |
| *E.coli* | Before | 0.787 | 0.778 | 0.796 | 0.575 | 302 | 86 | 79 | 309 |
| After | 0.832 | 0.858 | 0.807 | 0.666 | 333 | 55 | 75 | 313 |
| *G.subterruneus* | Before | 0.817 | 0.830 | 0.804 | 0.634 | 751 | 154 | 177 | 728 |
| After | 0.837 | 0.840 | 0.834 | 0.674 | 760 | 145 | 150 | 755 |
| *G.pickeringii* | Before | 0.832 | 0.838 | 0.826 | 0.664 | 477 | 92 | 99 | 470 |
| After | 0.860 | 0.863 | 0.858 | 0.721 | 491 | 78 | 81 | 488 |

**Table S3.** Top 5% most important features among the 700 features in six species.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | ***C.elegans*** | ***D.melanogaster*** | ***A.thaliana*** | ***E.coli*** | ***G.subterruneus*** | ***G.pickeringii*** |
| 1 | Feature109 | Feature184 | Feature184 | Feature636 | Feature169 | Feature169 |
| 2 | Feature637 | Feature607 | Feature85 | Feature633 | Feature183 | Feature185 |
| 3 | Feature640 | Feature610 | Feature626 | Feature87 | Feature173 | Feature173 |
| 4 | Feature184 | Feature96 | Feature623 | Feature105 | Feature180 | Feature190 |
| 5 | Feature628 | Feature99 | Feature248 | Feature224 | Feature185 | Feature174 |
| 6 | Feature631 | Feature173 | Feature180 | Feature270 | Feature224 | Feature186 |
| 7 | Feature246 | Feature85 | Feature244 | Feature580 | Feature201 | Feature180 |
| 8 | Feature638 | Feature77 | Feature106 | Feature577 | Feature104 | Feature587 |
| 9 | Feature641 | Feature87 | Feature99 | Feature626 | Feature186 | Feature590 |
| 10 | Feature98 | Feature623 | Feature232 | Feature623 | Feature72 | Feature206 |
| 11 | Feature103 | Feature626 | Feature641 | Feature480 | Feature250 | Feature189 |
| 12 | Feature207 | Feature89 | Feature638 | Feature600 | Feature313 | Feature201 |
| 13 | Feature248 | Feature248 | Feature612 | Feature597 | Feature174 | Feature575 |
| 14 | Feature105 | Feature170 | Feature615 | Feature169 | Feature232 | Feature572 |
| 15 | Feature625 | Feature76 | Feature607 | Feature74 | Feature228 | Feature702 |
| 16 | Feature622 | Feature244 | Feature610 | Feature587 | Feature189 | Feature77 |
| 17 | Feature94 | Feature183 | Feature96 | Feature590 | Feature407 | Feature205 |
| 18 | Feature112 | Feature612 | Feature93 | Feature178 | Feature249 | Feature249 |
| 19 | Feature99 | Feature615 | Feature200 | Feature72 | Feature231 | Feature250 |
| 20 | Feature175 | Feature180 | Feature488 | Feature593 | Feature68 | Feature170 |
| 21 | Feature117 | Feature592 | Feature440 | Feature596 | Feature243 | Feature441 |
| 22 | Feature186 | Feature595 | Feature636 | Feature611 | Feature73 | Feature202 |
| 23 | Feature111 | Feature93 | Feature633 | Feature608 | Feature205 | Feature228 |
| 24 | Feature200 | Feature232 | Feature100 | Feature80 | Feature416 | Feature270 |
| 25 | Feature182 | Feature247 | Feature183 | Feature206 | Feature247 | Feature72 |
| 26 | Feature108 | Feature488 | Feature247 | Feature249 | Feature408 | Feature515 |
| 27 | Feature190 | Feature597 | Feature175 | Feature242 | Feature614 | Feature512 |
| 28 | Feature250 | Feature600 | Feature89 | Feature96 | Feature117 | Feature527 |
| 29 | Feature254 | Feature189 | Feature222 | Feature176 | Feature244 | Feature530 |
| 30 | Feature86 | Feature504 | Feature210 | Feature185 | Feature102 | Feature223 |
| 31 | Feature174 | Feature92 | Feature350 | Feature660 | Feature424 | Feature60 |
| 32 | Feature502 | Feature74 | Feature170 | Feature657 | Feature480 | Feature662 |
| 33 | Feature335 | Feature202 | Feature611 | Feature186 | Feature441 | Feature665 |
| 34 | Feature104 | Feature205 | Feature608 | Feature77 | Feature601 | Feature377 |
| 35 | Feature101 | Feature200 | Feature617 | Feature183 | Feature215 | Feature175 |

**Table S4.** Performances of the SVM models and the other five machine learning algorithms.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Classifiers** | **Acc** | **SN** | **SP** | **MCC** | **TP** | **FN** | **FP** | **TN** |
| *C.elegans* | GBDT | 0.798 | 0.810 | 0.786 | 0.595 | 1258 | 296 | 333 | 1221 |
| KNN | 0.742 | 0.773 | 0.711 | 0.486 | 1202 | 352 | 449 | 1105 |
| LR | 0.791 | 0.802 | 0.781 | 0.582 | 1246 | 308 | 341 | 1213 |
| NB | 0.755 | 0.748 | 0.762 | 0.510 | 1163 | 391 | 370 | 1184 |
| RF | 0.737 | 0.708 | 0.765 | 0.474 | 1101 | 453 | 365 | 1189 |
| SVM | 0.815 | 0.824 | 0.807 | 0.631 | 1280 | 274 | 300 | 1254 |
| *D.melanogaster* | GBDT | 0.808 | 0.821 | 0.794 | 0.616 | 1453 | 316 | 364 | 1405 |
| KNN | 0.779 | 0.806 | 0.753 | 0.559 | 1425 | 344 | 437 | 1332 |
| LR | 0.815 | 0.828 | 0.801 | 0.629 | 1465 | 304 | 352 | 1417 |
| NB | 0.797 | 0.815 | 0.780 | 0.595 | 1442 | 327 | 390 | 1379 |
| RF | 0.767 | 0.784 | 0.751 | 0.535 | 1387 | 382 | 441 | 1328 |
| SVM | 0.830 | 0.838 | 0.822 | 0.661 | 1483 | 286 | 314 | 1455 |
| *A.thaliana* | GBDT | 0.772 | 0.776 | 0.768 | 0.545 | 1535 | 443 | 458 | 1520 |
| KNN | 0.728 | 0.736 | 0.719 | 0.455 | 1455 | 523 | 555 | 1423 |
| LR | 0.785 | 0.785 | 0.786 | 0.570 | 1552 | 426 | 424 | 1554 |
| NB | 0.750 | 0.751 | 0.748 | 0.499 | 1486 | 492 | 498 | 1480 |
| RF | 0.741 | 0.749 | 0.733 | 0.482 | 1481 | 497 | 528 | 1450 |
| SVM | 0.787 | 0.778 | 0.796 | 0.573 | 1538 | 440 | 404 | 1574 |
| *E.coli* | GBDT | 0.826 | 0.851 | 0.802 | 0.653 | 330 | 58 | 77 | 311 |
| KNN | 0.773 | 0.884 | 0.662 | 0.560 | 343 | 45 | 131 | 257 |
| LR | 0.818 | 0.838 | 0.799 | 0.637 | 325 | 63 | 78 | 310 |
| NB | 0.808 | 0.822 | 0.794 | 0.616 | 319 | 69 | 80 | 308 |
| RF | 0.742 | 0.773 | 0.711 | 0.485 | 300 | 88 | 112 | 276 |
| SVM | 0.832 | 0.858 | 0.807 | 0.666 | 333 | 55 | 75 | 313 |
| *G.subterruneus* | GBDT | 0.830 | 0.836 | 0.823 | 0.660 | 757 | 148 | 160 | 745 |
| KNN | 0.790 | 0.789 | 0.791 | 0.580 | 714 | 191 | 189 | 716 |
| LR | 0.826 | 0.836 | 0.815 | 0.652 | 757 | 148 | 167 | 738 |
| NB | 0.795 | 0.786 | 0.804 | 0.590 | 711 | 194 | 177 | 728 |
| RF | 0.767 | 0.759 | 0.775 | 0.534 | 687 | 218 | 204 | 701 |
| SVM | 0.837 | 0.840 | 0.834 | 0.674 | 760 | 145 | 150 | 755 |
| *G.pickeringii* | GBDT | 0.847 | 0.845 | 0.849 | 0.694 | 481 | 88 | 86 | 483 |
| KNN | 0.813 | 0.735 | 0.891 | 0.633 | 418 | 151 | 62 | 507 |
| LR | 0.840 | 0.824 | 0.856 | 0.680 | 469 | 100 | 82 | 487 |
| NB | 0.823 | 0.775 | 0.872 | 0.650 | 441 | 128 | 73 | 496 |
| RF | 0.793 | 0.793 | 0.793 | 0.585 | 451 | 118 | 118 | 451 |
| SVM | 0.860 | 0.863 | 0.858 | 0.721 | 491 | 78 | 81 | 488 |

**Table S5.** Performances of the proposed 4mcPred-SVM (this study) and two state-of-art predictors iDNA4mC and 4mcPred on six benchmark datasets from different species.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Predictors** | **Acc** | **SN** | **SP** | **MCC** | **TP** | **FN** | **FP** | **TN** |
| *C.elegans* | iDNA4mC | 0.786 | 0.797 | 0.775 | 0.572 | 1238 | 316 | 349 | 1205 |
| 4mCPred | 0.826 | 0.825 | 0.826 | 0.652 | 1282 | 272 | 270 | 1284 |
| This study | 0.815 | 0.824 | 0.807 | 0.631 | 1280 | 274 | 300 | 1254 |
| *D.melanogaster* | iDNA4mC | 0.812 | 0.833 | 0.791 | 0.625 | 1474 | 295 | 369 | 1400 |
| 4mCPred | 0.822 | 0.824 | 0.821 | 0.646 | 1458 | 311 | 317 | 1452 |
| This study | 0.830 | 0.838 | 0.822 | 0.661 | 1483 | 286 | 314 | 1455 |
| *A.thaliana* | iDNA4mC | 0.760 | 0.757 | 0.762 | 0.519 | 1498 | 480 | 471 | 1507 |
| 4mCPred | 0.768 | 0.755 | 0.780 | 0.536 | 1494 | 484 | 435 | 1543 |
| This study | 0.787 | 0.778 | 0.796 | 0.573 | 1538 | 440 | 404 | 1574 |
| *E.coli* | iDNA4mC | 0.799 | 0.820 | 0.778 | 0.598 | 318 | 70 | 86 | 302 |
| 4mCPred | 0.826 | 0.819 | 0.832 | 0.655 | 318 | 70 | 65 | 323 |
| This study | 0.833 | 0.858 | 0.807 | 0.666 | 333 | 55 | 75 | 313 |
| *G.subterruneus* | iDNA4mC | 0.815 | 0.822 | 0.808 | 0.630 | 745 | 161 | 174 | 732 |
| 4mCPred | 0.828 | 0.818 | 0.837 | 0.662 | 742 | 164 | 148 | 758 |
| This study | 0.837 | 0.840 | 0.834 | 0.674 | 760 | 145 | 150 | 755 |
| *G.pickeringii* | iDNA4mC | 0.831 | 0.824 | 0.838 | 0.663 | 469 | 100 | 92 | 477 |
| 4mCPred | 0.830 | 0.850 | 0.810 | 0.668 | 484 | 85 | 108 | 461 |
| This study | 0.860 | 0.863 | 0.858 | 0.721 | 491 | 78 | 81 | 488 |