

# YASS: similarity search in DNA sequences



Laurent Noé

LORIA/UHP Nancy, France

Gregory Kucherov

LORIA/INRIA Nancy, France

Corresponding email : laurent.noe@loria.fr



## Similarity search

- Identifying similarity regions in DNA sequences (local alignment) remains a fundamental problem in Bioinformatics. Detecting similarities is a necessary step in functional prediction, phylogenetic analysis, and many other biological studies.
- Exhaustive similarity search algorithms (Smith-Waterman<sup>[1]</sup>) take a prohibitive time on whole genome sequences. Most of heuristic algorithms are based on first searching for small exact repeats called *seeds* (by using suffix tree<sup>[7]</sup> or hashing techniques) that are then extended to larger similarity regions.

## Existing Tools

- FASTA<sup>[4]</sup> and BLAST<sup>[3]</sup> are complementary approaches, as each of them spends most of its time on doing what the other does quickly. FASTA spends its time on generating and sounTjng small seeds and its extension phase is trait-forward, whereas BLASTN spends less time to generate seeds (due to a bigger seed size) but tries to extend each one.
- Gapped-BLAST<sup>[3]</sup> introduces a two-hit criterion, which is a more selective, but less sensitive approach on small-score regions. PATTERN-HUNTER<sup>[2]</sup> proposes a more sensitive method using gapped seeds, and extends the two-hit criterion to allow an overlap between seeds.

## YASS Approach

We proposed YASS (*Yet Another Similarity Searcher*) – a new similarity search method (program available at [www.loria.fr/~noe](http://www.loria.fr/~noe)).

YASS algorithm is composed of two parts :

- chaining algorithm* links together *seeds* potentially belonging to the same similarity region.
- extension algorithm* triggers an extension of a group of seeds to a potential similarity region, according to an extension criterion.

## Chaining criteria

A *seed* is a pair of occurrences of the same k-mer. Two seeds are linked together if they verify two distance criteria :

- Inter-seed distance* (distance between the first or the second k-mers of the seeds) is below a given threshold, computed according to the *waiting time distribution* <sup>[1]</sup>.

- The variation between *intra-seed distances* (distances between the two k-mers of each seed) is below a given threshold, computed according to a *random walk distribution* <sup>[1]</sup>. This accounts for possible indels.

Both thresholds are estimated assuming a Bernoulli model of DNA sequence.

## Extension criterion

The extension of a group of (possibly overlapping) seeds is driven by the overall number of single nucleotide matches, called **group size**. Whenever the group size reaches a threshold, the extension is triggered.

## Other features

Low complexity regions can be filtered out, according to the triplet entropy.

The program outputs positions and an alignment (including score and e-value) of similarity regions. Other output information includes the *mutation bias of nucleotide triplets*, and *transitions/transversions* bias.

## Typical output

```

Terminal
File Edit Settings Help
...
180790 180791 180792 180793 180794 180795 180796 180797
...
180798 180799 180800 180801 180802 180803 180804 180805
...
180806 180807 180808 180809 180810 180811 180812 180813
...
180814 180815 180816 180817 180818 180819 180820 180821
...
180822 180823 180824 180825 180826 180827 180828 180829
...
180830 180831 180832 180833 180834 180835 180836 180837
...
180838 180839 180840 180841 180842 180843 180844 180845
...
180846 180847 180848 180849 180850 180851 180852 180853
...
180854 180855 180856 180857 180858 180859 180860 180861
...
180862 180863 180864 180865 180866 180867 180868 180869
...
180870 180871 180872 180873 180874 180875 180876 180877
...
180878 180879 180880 180881 180882 180883 180884 180885
...
180886 180887 180888 180889 180890 180891 180892 180893
...
180894 180895 180896 180897 180898 180899 180900 180901
...
180902 180903 180904 180905 180906 180907 180908 180909
...
180910 180911 180912 180913 180914 180915 180916 180917
...
180918 180919 180920 180921 180922 180923 180924 180925
...
180926 180927 180928 180929 180930 180931 180932 180933
...
180934 180935 180936 180937 180938 180939 180940 180941
...
180942 180943 180944 180945 180946 180947 180948 180949
...
180950 180951 180952 180953 180954 180955 180956 180957
...
180958 180959 180960 180961 180962 180963 180964 180965
...
180966 180967 180968 180969 180970 180971 180972 180973
...
180974 180975 180976 180977 180978 180979 180980 180981
...
180982 180983 180984 180985 180986 180987 180988 180989
...
180990 180991 180992 180993 180994 180995 180996 180997
...
180998 180999 190000 190001 190002 190003 190004 190005
...
190006 190007 190008 190009 190010 190011 190012 190013
...
190014 190015 190016 190017 190018 190019 190020 190021
...
190022 190023 190024 190025 190026 190027 190028 190029
...
190030 190031 190032 190033 190034 190035 190036 190037
...
190038 190039 190040 190041 190042 190043 190044 190045
...
190046 190047 190048 190049 190050 190051 190052 190053
...
190054 190055 190056 190057 190058 190059 190060 190061
...
190062 190063 190064 190065 190066 190067 190068 190069
...
190070 190071 190072 190073 190074 190075 190076 190077
...
190078 190079 190080 190081 190082 190083 190084 190085
...
190086 190087 190088 190089 190090 190091 190092 190093
...
190094 190095 190096 190097 190098 190099 190100 190101
...
190102 190103 190104 190105 190106 190107 190108 190109
...
190110 190111 190112 190113 190114 190115 190116 190117
...
190118 190119 190120 190121 190122 190123 190124 190125
...
190126 190127 190128 190129 190130 190131 190132 190133
...
190134 190135 190136 190137 190138 190139 190140 190141
...
190142 190143 190144 190145 190146 190147 190148 190149
...
190150 190151 190152 190153 190154 190155 190156 190157
...
190158 190159 190160 190161 190162 190163 190164 190165
...
190166 190167 190168 190169 190170 190171 190172 190173
...
190174 190175 190176 190177 190178 190179 190180 190181
...
190182 190183 190184 190185 190186 190187 190188 190189
...
190190 190191 190192 190193 190194 190195 190196 190197
...
190198 190199 190200 190201 190202 190203 190204 190205
...
190206 190207 190208 190209 190210 190211 190212 190213
...
190214 190215 190216 190217 190218 190219 190220 190221
...
190222 190223 190224 190225 190226 190227 190228 190229
...
190230 190231 190232 190233 190234 190235 190236 190237
...
190238 190239 190240 190241 190242 190243 190244 190245
...
190246 190247 190248 190249 190250 190251 190252 190253
...
190254 190255 190256 190257 190258 190259 190260 190261
...
190262 190263 190264 190265 190266 190267 190268 190269
...
190270 190271 190272 190273 190274 190275 190276 190277
...
190278 190279 190280 190281 190282 190283 190284 190285
...
190286 190287 190288 190289 190290 190291 190292 190293
...
190294 190295 190296 190297 190298 190299 190300 190301
...
190302 190303 190304 190305 190306 190307 190308 190309
...
190310 190311 190312 190313 190314 190315 190316 190317
...
190318 190319 190320 190321 190322 190323 190324 190325
...
190326 190327 190328 190329 190330 190331 190332 190333
...
190334 190335 190336 190337 190338 190339 190340 190341
...
190342 190343 190344 190345 190346 190347 190348 190349
...
190350 190351 190352 190353 190354 190355 190356 190357
...
190358 190359 190360 190361 190362 190363 190364 190365
...
190366 190367 190368 190369 190370 190371 190372 190373
...
190374 190375 190376 190377 190378 190379 190380 190381
...
190382 190383 190384 190385 190386 190387 190388 190389
...
190390 190391 190392 190393 190394 190395 190396 190397
...
190398 190399 190400 190401 190402 190403 190404 190405
...
190406 190407 190408 190409 190410 190411 190412 190413
...
190414 190415 190416 190417 190418 190419 190420 190421
...
190422 190423 190424 190425 190426 190427 190428 190429
...
190430 190431 190432 190433 190434 190435 190436 190437
...
190438 190439 190440 190441 190442 190443 190444 190445
...
190446 190447 190448 190449 190450 190451 190452 190453
...
190454 190455 190456 190457 190458 190459 190460 190461
...
190462 190463 190464 190465 190466 190467 190468 190469
...
190470 190471 190472 190473 190474 190475 190476 190477
...
190478 190479 190480 190481 190482 190483 190484 190485
...
190486 190487 190488 190489 190490 190491 190492 190493
...
190494 190495 190496 190497 190498 190499 190500 190501
...
190502 190503 190504 190505 190506 190507 190508 190509
...
190510 190511 190512 190513 190514 190515 190516 190517
...
190518 190519 190520 190521 190522 190523 190524 190525
...
190526 190527 190528 190529 190530 190531 190532 190533
...
190534 190535 190536 190537 190538 190539 190540 190541
...
190542 190543 190544 190545 190546 190547 190548 190549
...
190550 190551 190552 190553 190554 190555 190556 190557
...
190558 190559 190560 190561 190562 190563 190564 190565
...
190566 190567 190568 190569 190570 190571 190572 190573
...
190574 190575 190576 190577 190578 190579 190580 190581
...
190582 190583 190584 190585 190586 190587 190588 190589
...
190590 190591 190592 190593 190594 190595 190596 190597
...
190598 190599 190600 190601 190602 190603 190604 190605
...
190606 190607 190608 190609 190610 190611 190612 190613
...
190614 190615 190616 190617 190618 190619 190620 190621
...
190622 190623 190624 190625 190626 190627 190628 190629
...
190630 190631 190632 190633 190634 190635 190636 190637
...
190638 190639 190640 190641 190642 190643 190644 190645
...
190646 190647 190648 190649 190650 190651 190652 190653
...
190654 190655 190656 190657 190658 190659 190660 190661
...
190662 190663 190664 190665 190666 190667 190668 190669
...
190670 190671 190672 190673 190674 190675 190676 190677
...
190678 190679 190680 190681 190682 190683 190684 190685
...
190686 190687 190688 190689 190690 190691 190692 190693
...
190694 190695 190696 190697 190698 190699 190700 190701
...
190702 190703 190704 190705 190706 190707 190708 190709
...
190710 190711 190712 190713 190714 190715 190716 190717
...
190718 190719 190720 190721 190722 190723 190724 190725
...
190726 190727 190728 190729 190730 190731 190732 190733
...
190734 190735 190736 190737 190738 190739 190740 190741
...
190742 190743 190744 190745 190746 190747 190748 190749
...
190750 190751 190752 190753 190754 190755 190756 190757
...
190758 190759 190760 190761 190762 190763 190764 190765
...
190766 190767 190768 190769 190770 190771 190772 190773
...
190774 190775 190776 190777 190778 190779 190780 190781
...
190782 190783 190784 190785 190786 190787 190788 190789
...
190790 190791 190792 190793 190794 190795 190796 190797
...
190798 190799 190800 190801 190802 190803 190804 190805
...
190806 190807 190808 190809 190810 190811 190812 190813
...
190814 190815 190816 190817 190818 190819 190820 190821
...
190822 190823 190824 190825 190826 190827 190828 190829
...
190830 190831 190832 190833 190834 190835 190836 190837
...
190838 190839 190840 190841 190842 190843 190844 190845
...
190846 190847 190848 190849 190850 190851 190852 190853
...
190854 190855 190856 190857 190858 190859 190860 190861
...
190862 190863 190864 190865 190866 190867 190868 190869
...
190870 190871 190872 190873 190874 190875 190876 190877
...
190878 190879 190880 190881 190882 190883 190884 190885
...
190886 190887 190888 190889 190890 190891 190892 190893
...
190894 190895 190896 190897 190898 190899 190900 190901
...
190902 190903 190904 190905 190906 190907 190908 190909
...
190910 190911 190912 190913 190914 190915 190916 190917
...
190918 190919 190920 190921 190922 190923 190924 190925
...
190926 190927 190928 190929 190930 190931 190932 190933
...
190934 190935 190936 190937 190938 190939 190940 190941
...
190942 190943 190944 190945 190946 190947 190948 190949
...
190950 190951 190952 190953 190954 190955 190956 190957
...
190958 190959 190960 190961 190962 190963 190964 190965
...
190966 190967 190968 190969 190970 190971 190972 190973
...
190974 190975 190976 190977 190978 190979 190980 190981
...
190982 190983 190984 190985 190986 190987 190988 190989
...
190990 190991 190992 190993 190994 190995 190996 190997
...
190998 190999 191000 191001 191002 191003 191004 191005
...
191006 191007 191008 191009 191010 191011 191012 191013
...
191014 191015 191016 191017 191018 191019 191020 191021
...
191022 191023 191024 191025 191026 191027 191028 191029
...
191030 191031 191032 191033 191034 191035 191036 191037
...
191038 191039 191040 191041 191042 191043 191044 191045
...
191046 191047 191048 191049 191050 191051 191052 191053
...
191054 191055 191056 191057 191058 191059 191060 191061
...
191062 191063 191064 191065 191066 191067 191068 191069
...
191070 191071 191072 191073 191074 191075 191076 191077
...
191078 191079 191080 191081 191082 191083 191084 191085
...
191086 191087 191088 191089 191090 191091 191092 191093
...
191094 191095 191096 191097 191098 191099 191100 191101
...
191102 191103 191104 191105 191106 191107 191108 191109
...
191110 191111 191112 191113 191114 191115 191116 191117
...
191118 191119 191120 191121 191122 191123 191124 191125
...
191126 191127 191128 191129 191130 191131 191132 191133
...
191134 191135 191136 191137 191138 191139 191140 191141
...
191142 191143 191144 191145 191146 191147 191148 191149
...
191150 191151 191152 191153 191154 191155 191156 191157
...
191158 191159 191160 191161 191162 191163 191164 191165
...
191166 191167 191168 191169 191170 191171 191172 191173
...
191174 191175 191176 191177 191178 191179 191180 191181
...
191182 191183 191184 191185 191186 191187 191188 191189
...
191190 191191 191192 191193 191194 191195 191196 191197
...
191198 191199 191200 191201 191202 191203 191204 191205
...
191206 191207 191208 191209 191210 191211 191212 191213
...
191214 191215 191216 191217 191218 191219 191220 191221
...
191222 191223 191224 191225 191226 191227 191228 191229
...
191230 191231 191232 191233 191234 191235 191236 191237
...
191238 191239 191240
```