YASS: similarity search in DNA sequences

LaurenTNoé — 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[8]) 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[7] or hashing techniques) that are then extended to larger similarity regions.

Existing Tools

- FASTA[4] and BLAST[3] 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 cmall seeds and its extension phase is ŝtrait-forward, whereas BLASTN spends less time to generate seeds (due to a bigger seed size) but tries to extend each one.
- Gapped-BLAST[3] introduces a two-hit criterion, which is a more selective, but less sensitive approach on cmall-ssore regions. PATTERN-HUNTER[2] 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**).

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* [1].
- 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* [1]. This accounts for possible indels.

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

Extension criterion

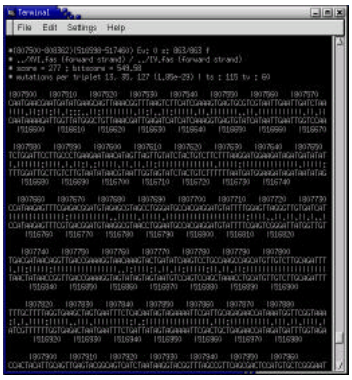
The extension of a group of (possibly overlapping) seeds is driven by the overall number of single nucleotide matches, called **group size**. Whenever the group size reaches a threshold, the extension is triggered.

Other features

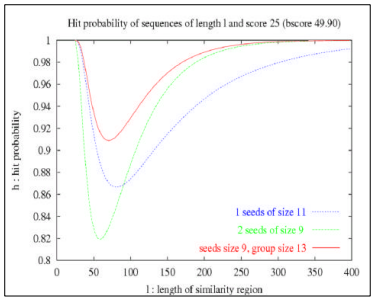
Low complexity regions can be filtered out, according to the triplet entropy.

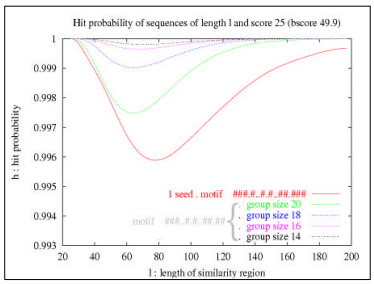
The program outputs positions and an alignment (including score and e-value) of similarity regions. Other output information includes the *mutation bias of nucleotide triplets* , and *transitions/ transversions* bias.

Typical output



Sensitivity





Running Time

size				cpu time		
seed	group	ñ	ä	t_{chain}	t_{align}	t_{total}
9	13	135	5	2s	3s	**5s**
9	11	135	5	2s	6s	**8s**
8	13	97	4	7s	6s	**13s**
8	11	97	4	7s	11s	**18s**
7	13	69	4	22s	35s	**57s**
7	11	69	4	22s	41s	**63s**

Running time of comparing chr V (576893bp) *vs.* chr IX (439907 bp) of *S.Cerevisiae* on a 2.Mhz Pentium IV Linux computer. Both main and complementary strands are compared.

Default score parameters of NCBI BLASTn are used. *Maximal scoring pairs* (MSP) are considered (no subalignement yields a bigger score).

Gapped Seeds

- Gapped seeds can improve the sensitivity/selectivity trade-off [2,5].
- The choice of specific gapped seeds can be tailored to a certain type of similarity regions. Bulher[4] (this conference proceedings) has proposed a tool called Mandala that finds heuristically the best gapped seed according to a mutational model represented by a boolean Markov chain.
- YASS is compatible with those approaches. To apply the group size extension criterion in the case of gapped seeds, YASS uses a finite automaton to update the group size when a new gapped seed has been identified, overlapping the previous one.

```
1111111111110111111
###..#.#..##.##
   ###..#.#..##.##
1111117117117117114
    ##..#.#..##.##   +3
1111111111117111111  17
```

Example

Alignment shown has been obtained by comparing *S.Cerevisiae* chromosomes IX and V. Note that it has only one contiguous seed of size 10.

- To find such an alignment with continuous seeds, one should use a smaller seed size, which leads to a time increase (13s for size 8).
- Another solution consists in using gapped seeds adapted for CDS. It is more efficient, both in sensitivity and selectivity, as the mutation bias per triplet is significant here.

According to the annotation of *S.Cerevisiae* (**http://pedant.gsf.de**), those regions are fragments of CDSs coding for two hypothetical proteins of the same family.

References:

[1] G.Benson, *Tandem repeats finder: a program to analyze DNA sequences* , Nucleic Acids Research, 1999,vol 27, 2, pp 573-580.
[2] B.Ma J.Tromp M.Li, *PatternHunter: Faster and more sensitive homology search* , Bioinformatics 2002, vol 18, 3, pp 440-445.
[3] S.Altschul and all. *G-BLAST and PSI-BLAST: a new generation of protein search programs* , Nucleic Acids Research, 1997,vol 25, 17 pp 3389-3402.
[4] D.Lipman W.Pearson, *Improved tools for biological sequence comparison* , Proc. Natl. Acad.Sci. USA, 1988, vol 85, pp 2444-2448.
[5] J.Bulher U.Keich Y.Sun, *Designing seeds for Similarity Search in Genomic DNA* , RECOMB 2003.
[6] S.Burkhardt J.Kärkkäinen, *Better Filtering with Gapped q-Grams*, Combinatorial Pattern Matching, 2001, pp 73-85
[7] E.Ukkonen *On-line Construction of Suffix-Trees*, Algorithmica 1995, vol 14, 249-260
[8] T. Smith M.Waterman, *Identification of common molecular subsequences* , Journal of Molecular Biology, 1981,vol 147, pp 195-197