

# YASS: similarity search in DNA sequences

Laurent Noé<sup>1</sup>, Grégory Kucherov<sup>2</sup>

**Keywords:** DNA, sequence repeat, local alignment, sequence similarity

## 1 Introduction.

Identifying similarity regions inside a DNA sequence (repeats), or between two sequences (local alignment), is a fundamental problem in bioinformatics. For this task, many algorithms use a technique based on searching for small exact repetitions of fixed size (*seeds*) and trying to extend those into larger approximate repeats. **BLAST** family [3] is the most prominent representative of this approach.

A slightly different but related method is implemented in **FASTA** [2] where the “hit strategy” is based on the count of the number of seeds found on each diagonal. **REPuter** [4] uses a different approach, based on suffix trees.

In **YASS** (Yet Another Similarity Searcher), we propose another method which tries to group together multiple seeds in order to consider a “distributed hit”, as in **FASTA**. But since we deal with larger seeds than **FASTA** (for the sake of speed), but smaller than **BLAST** (for the sake of sensitivity), we have formulated another hit criterion based on the overall seed size of a group (group size), which allows us to be more sensitive than **BL2SEQ** (Blast on two sequences) without loosing in time efficiency.

For example, if we consider approximate repeats scored 30 by **BLAST** (+1/-3), one will more frequently find a group of (possibly overlapping) seeds each of size at least 9 with group size at least 13, rather than one seed of size at least 11 (which is the default parameter of **BL2SEQ**).

Grouping multiple seeds reduces the number of infertile extensions, thus saving time for computing unnecessary alignment scores.

## 2 Method.

Parameters used to group seeds are estimated according to probability distributions, assuming a Bernoulli model of DNA sequence. Three probability criteria have been used:

- seed size and the minimal seed number which trigger grouping, are computed from the minimum repeat size and the minimal similarity rate between copies, according to the distribution of the longest run of one value in a Bernoulli sequence,
- maximal distance  $\rho$  between seeds inside one repeat copy is computed according to the *waiting time distribution*,
- indels are accounted for by computing a statistical bound  $\delta$  of *random walk distribution*, which simulates the variation of distance between corresponding seeds inside a repeat copy.

Similar criteria have been used in **Tandem Repeats Finder** [1] for computing criteria needed to find tandem repeats. The three criteria are used to *chain* seeds which potentially belong to the same approximate repeat. Chains found are then validated by computing a classical alignment score.

The chaining algorithm is linear on the number of seed pairs found.

---

<sup>1</sup>Loria/UHP Nancy France. E-mail: laurent.noe@loria.fr

<sup>2</sup>Loria/Inria Nancy France. E-mail: gregory.kucherov@loria.fr

### 3 Results and Discussion.

Tests have been carried out on chromosomes of *S.Cerevisiae* on a **Pentium IV 1.7 Ghz**. Running times are given below, depending on the *seed size* and *group size*.  $\delta$  and  $\rho$  have been introduced in Section 2. Comparison has been made with **BLAT**, **REPuter** as well as with **BL2SEQ**, in term of selectivity and sensitivity.

size				cpu time			results		
seed	group	$\delta$	$\rho$	$t_{chain}$	$t_{align}$	$t_{total}$	#chains	#extensions	#alignments
9	13	135	5	2s	2s	<b>4s</b>	759558	46546	5439
9	11	135	5	2s	6s	<b>8s</b>	759558	231955	58389
8	13	97	4	7s	7s	<b>14s</b>	2710181	205014	5755
8	11	97	4	7s	11s	<b>18s</b>	2710181	379643	56528
7	13	69	4	22s	35s	<b>57s</b>	10227901	1610229	9995
7	11	69	4	22s	41s	<b>69s</b>	10227901	1772701	62188

Table 1: Tests and Results on chromosomes V and chromosomes IX.

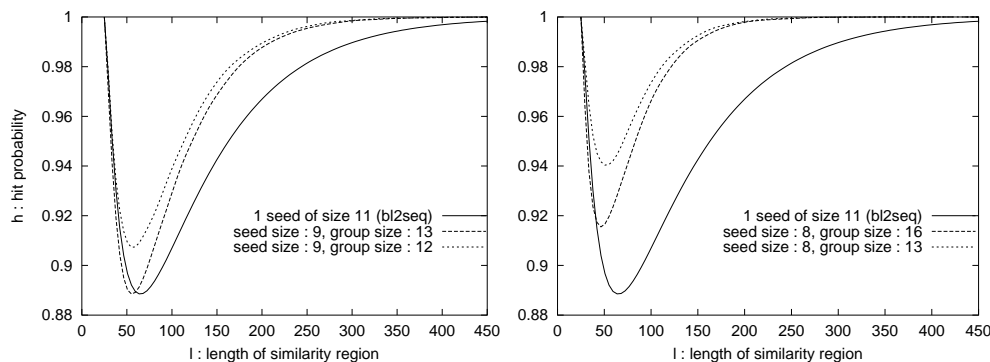


Figure 1: Hit probability on sequences of length  $l$  and score 25

Figure 1 illustrates the hit probability of our method compared to **BL2SEQ** for pairs of sequences scored 25. As a conclusion, our program runs in a time comparable to that of **BLAST**, but can be made more sensitive for detecting significant similarities.

### References

- [1] Benson, G. 1999. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Research*, 27(2):573–580.
- [2] Lipman, D. and Pearson, W. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448.
- [3] Altschul, S. and Madden, T. and Schäffer, A. and Zhang, J. and Zhang, Z. and Miller, W. and Lipman, D. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402
- [4] Kurtz, S. and Ohlebusch, E. and Schleiermacher, C. and Stoye, J. 2001 REPuter: the manifold applications of repeat analysis *Nucleic Acids Research*, 29(22):4633–4642