

A new method of finding similarity regions in DNA sequences

Laurent NOE, Gregory KUCHEROV

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

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 [1] is the most prominent representative of this approach. ASSIRC [7] is another example. A slightly different but related method is implemented in FASTA [6]. REPuter [5] and MUMmer [4] use a different approach, based on suffix trees.

We propose a new method which tries to group together multiple seeds, in order to form rapidly large similarity regions instead of extending individual seeds. In a very restricted form, this idea has been used in late versions of Blast [2]. Here we push it much further, and come up with a more sensitive approach, allowing for smaller seed sizes without considerable drop in time efficiency. For example, if we consider approximate repeats of size at least 100 with 75% of similarity between copies, one finds more frequently 3 (or more) distinct seeds of size at least 7 than one (or more) seed of size at least 11 (which is the default parameter of BLASTN). Grouping multiple seeds also reduces the number of infertile extensions, thus saving time for computing unnecessary alignment scores.

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:

- the seed size and the minimal seed number which triggers 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,
- the maximal distance ρ between seeds inside one repeat copy is computed according to *waiting time distribution*,
- indels are accounted for by computing a statistical bound δ of *random walk distribution*, which simulates the variation of distance between corresponding seeds inside a repeat copy.

Similar criteria have been used in [3] for computing 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. Running time of test runs is given below.

Results and Discussion

Tests have been carried out on chromosomes V and IX of *S. Cerevisiae*¹ (respectively 576869 and 439885 bps) on a Pentium III 1Ghz.

Compared to BLAST, our algorithm runs in a comparable time, but appeared to be more sensitive in finding repetitions with good similarity level (more than 70%) between copies. For example, the following repeats have been missed by BLAST but found by our program:

¹available through <http://genome-www4.stanford.edu/cgi-bin/SGD/seqTools>

k	ρ	δ	# of alignments	# of reps found	time
9	85	10	212	60	3 s
8	64	9	701	64	8 s
7	47	7	748	68	27 s
6	35	6	2388	72	100 s
5	25	5	7863	72	425 s

Table 1: Results on chromosomes V and IX of *S.Cerevisiae*

V.fas	IX.fas	size	BLAST score
270757–271078	257864–258185	321	836
273624–273841	259542–259759	217	498
435232–435429	336678–336868	197/190	461
449152–449284	336729–336862	132/133	338

Table 2: Repeats characterized by small seeds

On the other hand, our algorithm is much faster than ASSIRC [7] for a similar input. Comparisons with REPuter [5] have shown that it tends to keep apart fragments that form a longer repetition. In contrast, our program assembles fragments in a flexible way, allowing for indels of size smaller than δ .

References

- [1] Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [2] Stephen Altschul, Thomas Madden, Alejandro Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
- [3] Gary Benson. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.
- [4] Arthur Delcher, Simon Kasif, Robert Fleischmann, Jeremy Peterson, Owen White, and Salzberg Steven. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.
- [5] Stefan Kurtz and Chris Schleiermacher. Reputer: fast computation of maximal repeats in complete genome. *Bioinformatics*, 15(5):426–427, 1999.
- [6] William Pearson and David Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.
- [7] Pierre Vincens, Laurent Buffat, Cécile André, Jean-Paul Chevrolat, Jean-Francois Boisvieux, and Serge Hazout. A strategy for finding regions of similarity in complete genome sequences. *Bioinformatics*, 14(8):715–725, 1998.