REPuter: Fast Computation of Maximal Repeats in Complete Genomes

Stefan Kurtz and Chris Schleiermacher

Technische Fakultät, Universität Bielefeld
Postfach 10 01 31, D-33501 Bielefeld, Germany
Email: {kurtz,icschlei}@techfak.uni-bielefeld.de

Abstract for Software Demonstration

We demonstrate a software tool *REPuter* which allows to determine all exact repetitive substrings contained in complete genomes. Exact repeats are only a small fraction of all repeats of biological interest. However, since exact repeats usually form the core blocks of approximate repeats, *REPuter* can also serve as a fast subroutine for programs that detect these. The running time and space requirement of *REPuter* is linear in the length of the genome and in the size of the output. The method we use is thus asymtotically optimal. The main advantage over previous programs is the reduced time and space complexity of our tool, and the guaranteed linear worst case behavior. This allows to handle much longer input sequences. The current version is able to process input sequences consisting of up to 67 million bases. The following table shows time and space requirements for computing exact **F**orward repeats and **P**alindromic repeats of length ≥ 20 for different genomes (SUN Enterprise, 166 MHz, 512 MB RAM, and Pentium II, 350 MHz, 384 MB RAM).

Input, approx.	**F**-Repeats				**P**-Repeats			
size (MB)	Num. of repeats	SUN (s)	PII (s)	Mem. (MB)	Num. of repeats	SUN (s)	PII (s)	Mem. (MB)
M. gen , 0.6	1022	4	2	8	56	9	4	15
C. tra , 1.0	26	8	3	14	8	17	7	27
A. ful , 2.1	4197	18	7	29	3267	39	16	57
E. coli, 4.6	7799	40	17	61	6763	83	35	122
S. cer, 11.58	173526	112	46	160	171841	248	97	320

REPuter consists of two programs, a search engine and a visualizing component. The search engine processes a DNA sequence given by the user in Fasta-format, and returns a representation of all maximal repeats in a simple ASCII-format. The visualizing component processes the output of the search engine and produces an overview of the number, the length, and the location of the repeated substrings. Both components will be demonstrated.

For more information on *REPuter*, see the Bielefeld Bioinformatics Server `http://bibiserv.techfak.uni-bielefeld.de/reputer/` and the following paper:

[1] Kurtz, S. and Schleiermacher, C., *REPuter: Fast Computation of Maximal Repeats in Complete Genomes*, Bioinformatics, in press, 1999