

Name	$\sigma$	$n/z$	$n/2^{20}$	Source	Description
dna	16	14.2	100	S	Human genome
english	215	14.1	100	S	Gutenberg Project
sources	227	16.8	100	S	Linux and GCC sources
cere	5	84	100	R	yeast genome
einstein	121	2947	100	R	Wikipedia articles
kernel	160	156	100	R	Linux Kernel sources

Table 1: Data set used in the experiments. The files are 100MB prefixes of files from the Pizza & Chili standard corpus<sup>2</sup> (S) and the Pizza & Chili repetitive corpus<sup>3</sup> (R). The value of  $n/z$  (average length of an LZ77 phrase) is included as a measure of repetitiveness.