Chapter 3

A Repetitive Corpus Testbed

In this chapter we present a corpus of repetitive texts. These texts are categorized according to the source they come from into the following: Artificial Texts, Pseudo-Real Texts and Real Texts. The main goal of this collection is to serve as a standard testbed for benchmarking algorithms oriented to repetitive texts. The corpus can be downloaded from `http://pizzachili.dcc.uchile.cl/repcorpus.html`.

3.1 Artificial Texts

This subset is composed of highly repetitive texts that do not come from any real-life source, but are artificially generated through some mathematical definition and have interesting combinatorial properties.

3.1.1 Fibonacci Sequence (F_n)

This sequence is defined by the recurrence

$$\begin{aligned} F_1 &= \mathtt{0} \\ F_2 &= \mathtt{1} \\ F_n &= F_{n-1}F_{n-2} \end{aligned} \tag{3.1}$$

The length of the string F_n is the Fibonacci number f_n and the sequence is a *sturmian word* [Lot02], which means it has $i+1$ different substrings (factors) of length i.

3.1.2 Thue-Morse Sequence (T_n)

This sequence [AS99] is defined by the recurrence

$$T_1 = \texttt{0}$$
$$T_n = T_{n-1}\overline{T_{n-1}} \tag{3.2}$$

where $\bar{F}$ is the bitwise negation operator (i.e., all `0` get converted to `1` and all `1` to `0`). Because of the construction scheme of this sequence, there are many substrings of the form XX, for any string X. However, there are no overlapping squares, i.e., substrings of the form $\texttt{0}X\texttt{0}X\texttt{0}$ or $\texttt{1}X\texttt{1}X\texttt{1}$. Furthermore, this sequence is strongly cube-free, i.e., there are no substrings of the form XXx, where x is the first character of the string X. Another interesting property of this string is that it is recurrent. That is, given any finite substring w of length n, there is some length n_w (often much longer than n) such that w is contained in every substring of length n_w. The length of these strings is $|T_n| = 2^n$.

3.1.3 Run-Rich String Sequence (R_n)

A measure of string complexity, related to the regularities of the text and strongly related to the LZ77 parsing [KK99], is the number of runs.

Definition 3.1 ([Mai89])**.** The substring $T[i, j]$ is a *run* in a string T iff the minimum period of $T[i, j] = p \leq |T[i, j]|/2$ and $T[i, j]$ is not extendable to the right ($j = n$ or $T[j + 1] \neq T[j - p + 1]$) or left ($i = 1$ or $T[i - 1] \neq T[i + p - 1]$).

The higher the number of runs in a string, the more regularities it has.

It has been shown that the maximum number of runs in a string is greater than $0.944n$ [MKI+08] and lower than $1.029n$ [CIT08]. Franek *et al.* [FSS03] show a constructive and simple way to obtain strings with many runs; the n-th of those strings is denoted R_n. The ratio of the runs of their strings to the length approaches $3/(1 + \sqrt{5}) = 0.92705\ldots$.

3.2 Pseudo-Real Texts

Here we present a set of texts that were generated by artificially adding repetitiveness to real texts, thus we call them *pseudo-real texts*.

To generate the texts, we took a prefix of 1MiB of all texts of Pizza&Chili Corpus[1] and we mutated them. Our mutations take a random character position and change it to a random character different from the original one.

We used two different schemes for the mutations. The first one, denoted by a '1', generates different mutations of the first text. The second, denoted by a '2', mutates the last text generated. The second scheme resembles the changes obtained through time in a software project or the versions of a document.

The mutation rate, i.e., percentage of mutated characters, was set to 0.1%, 0.01% and 0.001%.

The base texts (all from the PizzaChili corpus) we mutated were the following:

- Sources: This file is formed by C/Java source code obtained by concatenating all the `.c`, `.h`, `.C` and `.java` files of the linux-2.6.11.6 and gcc-4.0.0 distributions.
- Pitches: This file is a sequence of midi pitch values (bytes in 0-127, plus a few extra special values) obtained from a myriad of MIDI files freely available on Internet.
- Proteins: This file is a sequence of newline-separated protein sequences obtained from the Swissprot database.
- DNA: This file is a sequence of newline-separated gene DNA sequences obtained from files `01hgp10` to `21hgp10`, plus `0xhgp10` and `0yhgp10`, from Gutenberg Project.
- English: This file is the concatenation of English text files selected from `etext02` to `etext05` collections of Gutenberg Project.
- XML: This file is an XML that provides bibliographic information on major computer science journals and proceedings and it was obtained from `http://dblp.uni-trier.de`.

3.3 Real Texts

This subset is composed of texts coming from real repetitive sources. These sources are DNA, Wikipedia Articles, Source Code, and Documents.

[1] `http://pizzachili.dcc.uchile.cl`

For the case of DNA we concatenated the texts in random order. For the others, we concatenated the texts according to the date they were created, from oldest to newest.

3.3.1 DNA

Our DNA texts come from different sources.

- The Saccharomyces Genome Resequencing Project[2] provides two text collections: *para*, which contains 36 sequences of *Saccharomyces Paradoxus* and *cere*, which contains 37 sequences of *Saccharomyces Cerevisiae*.
- From the National Center for Biotechnology Information (NCBI)[3] we collected some DNA sequences of the same bacteria. The species we collected are *Escherichia Coli* (23), *Salmonella Enterica* (15), *Staphylococcus Aureus* (14), *Streptococcus Pyogenes* (13), *Streptococcus Pneumoniae* (11) and *Clostridium Botulium* (10). We chose these species as they were the only ones with 10 or more different sequences.
- A collection composed of 78,041 sequences of *Haemophilus Influenzae*[4], also coming from the NCBI.

Remark 3.2. Although there are four bases $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, DNA sequences may have alphabets of size up to $16 = 2^4$ because some characters denote an unknown choice among the four bases. The most common character used is `N`, which denotes a totally unknown symbol.

3.3.2 Wikipedia Articles

We downloaded all versions of three Wikipedia articles, *Albert Einstein*, *Alan Turing* and *Nobel Prize*. We downloaded them in English (denoted *en*) and German (denoted *de*). We chose these languages as they are among the most widely used on Internet and their alphabet may be represented using standard 1-byte encodings. The versions for all documents are up to January 12, 2010, except for the English article of *Albert Einstein*, which was downloaded only up to November 10, 2006 because of the massive number of versions it has.

[2] `http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp`
[3] `http://www.ncbi.nlm.nih.gov`
[4] `ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/influenza.fna.gz`

3.3.3 Source Code

We collected all versions 5.x of the *Coreutils*[5] package and removed all binary files, making a total of 9 versions. We also collected all 1.0.x and 1.1.x versions of the *Linux Kernel*[6], making a total of 36 versions.

3.3.4 Documents

We took all *pdf* files of CIA World Leaders[7] from January 2003 to December 2009, and converted them to text (using software `pdftotext`).

3.4 Statistics

To understand the characteristics of the texts present in the *Repetitive Corpus*, we provide below some statistics about them. The statistics presented are the following:

- **Alphabet Size:** We give the alphabet size and the inverse probability of matching (IPM), which is the inverse of the probability that two characters chosen at random match. IPM is a measure of the effective alphabet size. On a uniformly distributed text, it is precisely the alphabet size.
- **Compression Ratio:** Since we are dealing with compressed indexes it is useful to have an idea of the compressibility of the texts using general-purpose compressors. The following compressors are used: `gzip`[8] gives an idea of compressibility via dictionaries (an LZ77 parsing with limited window size); `bzip2`[9] gives an idea of block-sorting compressibility (using the BWT transform, Section 2.12); `ppmdi`[10] gives an idea of partial-match-based compressors (related to the k-th order entropy, Section 2.3); `p7zip`[11] gives an idea of LZ77 compression with virtually unlimited window; and `Re-Pair`[12] [LM00] gives an idea of

[5] `ftp://mirrors.kernel.org/gnu/coreutils`
[6] `ftp://ftp.kernel.org/pub/linux/kernel`
[7] `https://www.cia.gov/library/publications/world-leaders-1`
[8] `http://www.gzip.org`
[9] `http://www.bzip.org`
[10] `http://pizzachili.dcc.uchile.cl/utils/ppmdi.tar.gz`
[11] `http://www.7-zip.org`
[12] `http://www.cbrc.jp/~rwan/en/restore.html`

grammar-based compression. All compressors were run with the highest compression options.

- **Empirical Entropy:** Here we give the empirical entropy H_k of the text with k ranging from 0 to 8, measured as compression ratio. We also show, in parentheses, the *complexity function* of T [Lot02] (or the *number of contexts*) which count how many different substrings of a given size does T have. This is exactly our $C(T,k)$ of Lemma 2.12. This measure has the following properties:

$$\begin{aligned} C(T,1) &= \sigma \\ C(T,n+m) &\leq C(T,n)C(T,m) \end{aligned}$$

  The lower this measure, the more repetitive the text is. For example, if $C(T,n) = 1\,\forall n$, then $T = c^m$ for some character c. When $P(C,n) = n+1$ the sequence is said to be *Sturmian* (the Fibonacci sequence is an example of a *Sturmian* string).

Remark 3.3. The compression ratios are given as the percentage of the compressed file size over the uncompressed file size, assuming the original file uses one byte per character. This means that 25% compression can be achieved over a DNA sequence having an alphabet {A,C,G,T} by simply using 2 bits per symbol. As seen from the real-life examples given, these four symbols are usually predominant, so it is not hard to get very close to 25% on general DNA sequences as well.

3.4.1 Artificial Texts

Tables 3.1-3.3 give the statistics of artificial texts.

File	Size	Σ	IPM
F_{41}	256MiB	2	1.894
T_{29}	256MiB	2	2.000
R_{13}	207MiB	2	2.000

Table 3.1: Alphabet statistics for Artificial Collection

File	**p7zip**	**bzip2**	**gzip**	**ppmdi**	**Re-pair**
F_{41}	0.17624%	0.00572%	0.46875%	1.87500%	0.00002%
T_{29}	0.35896%	0.01259%	0.54688%	2.18750%	0.00004%
R_{13}	0.17172%	0.01227%	0.53140%	2.12560%	0.00009%

Table 3.2: Compression statistics for Artificial Collection

File	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
F_{41}	11.99% (1)	7.41% (2)	4.58% (3)	4.58% (4)	2.83% (5)	2.83% (6)	2.83% (7)	1.75% (8)	1.75% (9)
T_{29}	12.50% (1)	11.48% (2)	8.34% (4)	8.34% (6)	4.16% (10)	4.16% (12)	4.16% (16)	2.09% (20)	2.09% (22)
R_{13}	12.50% (1)	9.85% (2)	8.51% (4)	6.55% (6)	2.56% (8)	2.33% (10)	2.33% (12)	2.33% (14)	2.33% (16)

Table 3.3: Empirical entropy statistics for Artificial Collection

3.4.2 Pseudo-Real Texts

Tables 3.4-3.9 give the statistics of pseudo-real texts.

File	Size	Σ	IPM
Xml 0.001%[1]	100MiB	89	27.84
Xml 0.01%[1]	100MiB	89	27.84
Xml 0.1%[1]	100MiB	89	27.84
DNA 0.001%[1]	100MiB	5	3.98
DNA 0.01%[1]	100MiB	5	3.98
DNA 0.1%[1]	100MiB	5	3.98
English 0.001%[1]	100MiB	106	15.65
English 0.01%[1]	100MiB	106	15.65
English 0.1%[1]	100MiB	106	15.65
Pitches 0.001%[1]	100MiB	73	33.07
Pitches 0.01%[1]	100MiB	73	33.07
Pitches 0.1%[1]	100MiB	73	33.07
Proteins 0.001%[1]	100MiB	21	16.90
Proteins 0.01%[1]	100MiB	21	16.90
Proteins 0.1%[1]	100MiB	21	16.90
Sources 0.001%[1]	100MiB	98	28.86
Sources 0.01%[1]	100MiB	98	28.86
Sources 0.1%[1]	100MiB	98	28.86

Table 3.4: Alphabet statistics for Pseudo-Real Collection (Scheme 1)

File	Size	Σ	IPM
Xml 0.001%[2]	100MiB	89	27.84
Xml 0.01%[2]	100MiB	89	27.84
Xml 0.1%[2]	100MiB	89	27.86
DNA 0.001%[2]	100MiB	5	3.98
DNA 0.01%[2]	100MiB	5	3.98
DNA 0.1%[2]	100MiB	5	3.98
English 0.001%[2]	100MiB	106	15.65
English 0.01%[2]	100MiB	106	15.66
English 0.1%[2]	100MiB	106	15.74
Pitches 0.001%[2]	100MiB	73	33.07
Pitches 0.01%[2]	100MiB	73	33.07
Pitches 0.1%[2]	100MiB	73	33.10
Proteins 0.001%[2]	100MiB	21	16.90
Proteins 0.01%[2]	100MiB	21	16.90
Proteins 0.1%[2]	100MiB	21	16.92
Sources 0.001%[2]	100MiB	98	28.86
Sources 0.01%[2]	100MiB	98	28.86
Sources 0.1%[2]	100MiB	98	28.92

Table 3.5: Alphabet statistics for Pseudo-Real Collection (Scheme 2)

File	p7zip	bzip2	gzip	ppmdi	Re-Pair
Xml 0.001%[1]	0.15%	11.00%	18.00%	3.50%	0.19%
Xml 0.01%[1]	0.18%	12.00%	18.00%	3.60%	0.46%
Xml 0.1%[1]	0.46%	12.00%	18.00%	4.10%	2.00%
DNA 0.001%[1]	0.27%	27.00%	28.00%	11.00%	0.34%
DNA 0.01%[1]	0.29%	27.00%	28.00%	11.00%	0.58%
DNA 0.1%[1]	0.51%	27.00%	28.00%	12.00%	2.50%
English 0.001%[1]	0.31%	28.00%	37.00%	22.00%	0.39%
English 0.01%[1]	0.35%	28.00%	37.00%	22.00%	0.65%
English 0.1%[1]	0.59%	28.00%	37.00%	22.00%	2.70%
Pitches 0.001%[1]	0.47%	54.00%	52.00%	47.00%	0.69%
Pitches 0.01%[1]	0.50%	54.00%	52.00%	47.00%	0.95%
Pitches 0.1%[1]	0.75%	54.00%	52.00%	48.00%	3.20%
Proteins 0.001%[1]	0.32%	41.00%	39.00%	31.00%	0.42%
Proteins 0.01%[1]	0.35%	41.00%	39.00%	31.00%	0.68%
Proteins 0.1%[1]	0.59%	41.00%	39.00%	32.00%	2.70%
Sources 0.001%[1]	0.20%	19.00%	25.00%	12.00%	0.28%
Sources 0.01%[1]	0.23%	19.00%	25.00%	12.00%	0.56%
Sources 0.1%[1]	0.50%	20.00%	25.00%	13.00%	2.60%

Table 3.6: Compression statistics for Pseudo-Real Collection (Scheme 1)

File	p7zip	bzip2	gzip	ppmdi	Re-Pair
Xml 0.001%[2]	0.15%	12.00%	18.00%	3.50%	0.18%
Xml 0.01%[2]	0.18%	14.00%	19.00%	4.40%	0.39%
Xml 0.1%[2]	0.39%	25.00%	29.00%	17.00%	2.20%
DNA 0.001%[2]	0.26%	27.00%	28.00%	11.00%	0.33%
DNA 0.01%[2]	0.29%	27.00%	28.00%	11.00%	0.52%
DNA 0.1%[2]	0.46%	27.00%	28.00%	13.00%	2.20%
English 0.001%[2]	0.31%	28.00%	37.00%	22.00%	0.38%
English 0.01%[2]	0.34%	29.00%	37.00%	23.00%	0.59%
English 0.1%[2]	0.55%	38.00%	43.00%	31.00%	2.50%
Pitches 0.001%[2]	0.46%	54.00%	52.00%	47.00%	0.68%
Pitches 0.01%[2]	0.49%	54.00%	53.00%	48.00%	0.89%
Pitches 0.1%[2]	0.71%	59.00%	57.00%	52.00%	2.80%
Proteins 0.001%[2]	0.31%	41.00%	39.00%	32.00%	0.41%
Proteins 0.01%[2]	0.34%	42.00%	40.00%	33.00%	0.62%
Proteins 0.1%[2]	0.54%	47.00%	46.00%	40.00%	2.50%
Sources 0.001%[2]	0.20%	20.00%	25.00%	13.00%	0.27%
Sources 0.01%[2]	0.23%	21.00%	26.00%	14.00%	0.49%
Sources 0.1%[2]	0.44%	34.00%	35.00%	26.00%	2.50%

Table 3.7: Compression statistics for Pseudo-Real Collection (Scheme 2)

File	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
Xml	65.25%	38.63%	21.00%	12.50%	8.13%	6.00%	5.25%	4.75%	4.13%
0.001%[1]	(1)	(89)	(3325)	(20560)	(56120)	(98084)	(134897)	(168846)	(200451)
Xml	65.25%	38.63%	21.00%	12.50%	8.13%	6.00%	5.25%	4.75%	4.13%
0.01%[1]	(1)	(89)	(4135)	(30975)	(79379)	(131811)	(177924)	(220923)	(261651)
Xml	65.25%	38.75%	21.25%	12.75%	8.25%	6.13%	5.38%	4.88%	4.25%
0.1%[1]	(1)	(89)	(5251)	(67479)	(196554)	(326296)	(440199)	(550570)	(661284)
DNA	25.00%	24.25%	24.13%	24.00%	24.00%	23.75%	23.50%	22.88%	21.25%
0.001%[1]	(1)	(5)	(18)	(67)	(260)	(1029)	(4102)	(16349)	(62437)
DNA	25.00%	24.25%	24.13%	24.00%	24.00%	23.75%	23.50%	22.88%	21.25%
0.01%[1]	(1)	(5)	(18)	(67)	(260)	(1029)	(4102)	(16368)	(63204)
DNA	25.00%	24.25%	24.13%	24.00%	24.00%	23.75%	23.50%	22.88%	21.38%
0.1%[1]	(1)	(5)	(19)	(70)	(264)	(1034)	(4109)	(16399)	(65168)
English	57.25%	45.13%	34.75%	25.88%	19.88%	15.88%	12.50%	9.63%	7.25%
0.001%[1]	(1)	(106)	(2659)	(18352)	(63299)	(145194)	(256838)	(379514)	(501400)
English	57.25%	45.13%	34.75%	25.88%	19.88%	15.88%	12.50%	9.63%	7.25%
0.01%[1]	(1)	(106)	(3243)	(24063)	(82896)	(180401)	(305292)	(439387)	(572056)
English	57.25%	45.25%	34.88%	26.13%	20.13%	16.00%	12.50%	9.75%	7.25%
0.1%[1]	(1)	(106)	(4491)	(46116)	(190765)	(439130)	(715127)	(983435)	(1237512)
Pitches	66.13%	61.00%	53.50%	37.13%	16.38%	6.25%	2.88%	1.38%	0.75%
0.001%[1]	(1)	(73)	(3549)	(73664)	(376958)	(642406)	(767028)	(833456)	(871970)
Pitches	66.13%	61.00%	53.50%	37.25%	16.38%	6.25%	2.88%	1.38%	0.75%
0.01%[1]	(1)	(73)	(3581)	(76900)	(399435)	(684445)	(821533)	(898126)	(946219)
Pitches	66.13%	61.13%	53.63%	37.38%	16.63%	6.38%	2.88%	1.50%	0.88%
0.1%[1]	(1)	(73)	(3733)	(95838)	(598394)	(1096014)	(1363610)	(1543086)	(1687166)
Proteins	52.25%	52.13%	51.63%	47.50%	25.13%	4.63%	0.75%	0.25%	0.25%
0.001%[1]	(1)	(21)	(422)	(8045)	(128975)	(463357)	(572530)	(589356)	(595906)
Proteins	52.25%	52.13%	51.63%	47.50%	25.13%	4.63%	0.75%	0.25%	0.25%
0.01%[1]	(1)	(21)	(422)	(8045)	(131064)	(494845)	(626269)	(654067)	(670075)
Proteins	52.25%	52.13%	51.63%	47.50%	25.50%	4.88%	0.88%	0.38%	0.38%
0.1%[1]	(1)	(21)	(425)	(8076)	(143879)	(768510)	(1150595)	(1293347)	(1403589)
Sources	68.75%	46.88%	30.00%	19.63%	14.38%	11.00%	8.38%	6.88%	5.75%
0.001%[1]	(1)	(98)	(4557)	(29667)	(75316)	(130527)	(194105)	(259413)	(320468)
Sources	68.75%	46.88%	30.00%	19.63%	14.38%	11.00%	8.50%	6.88%	5.75%
0.01%[1]	(1)	(98)	(5621)	(42303)	(102977)	(170525)	(244755)	(320237)	(391260)
Sources	68.75%	47.00%	30.25%	19.88%	14.63%	11.13%	8.50%	7.00%	5.88%
0.1%[1]	(1)	(98)	(7359)	(104679)	(299799)	(498046)	(687941)	(872189)	(1049051)

Table 3.8: Empirical entropy statistics for Pseudo-Real Collection (Scheme 1)

File	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
Xml	65.25%	38.63%	21.13%	12.63%	8.13%	6.00%	5.25%	4.75%	4.13%
$0.001\%^2$	(1)	(89)	(3325)	(20560)	(56120)	(98084)	(134897)	(168846)	(200451)
Xml	65.25%	39.38%	22.00%	13.25%	8.63%	6.50%	5.63%	5.13%	4.50%
$0.01\%^2$	(1)	(89)	(4135)	(31042)	(79630)	(132163)	(178388)	(221499)	(262329)
Xml	65.25%	44.00%	28.75%	18.50%	12.25%	9.25%	8.00%	7.13%	6.25%
$0.1\%^2$	(1)	(89)	(5255)	(72227)	(226418)	(378994)	(513539)	(645141)	(777226)
DNA	25.00%	24.25%	24.13%	24.00%	24.00%	23.75%	23.50%	22.88%	21.25%
$0.001\%^2$	(1)	(5)	(18)	(67)	(260)	(1029)	(4102)	(16349)	(62436)
DNA	25.00%	24.25%	24.13%	24.13%	24.00%	23.88%	23.50%	23.00%	21.38%
$0.01\%^2$	(1)	(5)	(18)	(67)	(260)	(1029)	(4102)	(16369)	(63242)
DNA	25.00%	24.50%	24.38%	24.25%	24.25%	24.13%	23.88%	23.50%	22.38%
$0.1\%^2$	(1)	(5)	(19)	(70)	(264)	(1034)	(4109)	(16400)	(65387)
English	57.25%	45.13%	34.75%	26.00%	20.00%	15.88%	12.50%	9.63%	7.13%
$0.001\%^2$	(1)	(106)	(2659)	(18353)	(63300)	(145195)	(256838)	(379514)	(501400)
English	57.25%	45.50%	35.38%	26.50%	20.25%	15.88%	12.38%	9.50%	7.13%
$0.01\%^2$	(1)	(106)	(3243)	(24079)	(83037)	(180592)	(305458)	(439539)	(572186)
English	57.38%	47.75%	39.50%	31.13%	23.00%	16.63%	12.13%	8.88%	6.38%
$0.1\%^2$	(1)	(106)	(4482)	(47357)	(202366)	(466838)	(749065)	(1015587)	(1265447)
Pitches	66.13%	61.13%	53.63%	37.25%	16.38%	6.25%	2.88%	1.38%	0.75%
$0.001\%^2$	(1)	(73)	(3549)	(73664)	(376958)	(642406)	(767028)	(833456)	(871970)
Pitches	66.13%	61.13%	53.88%	37.50%	16.50%	6.38%	2.88%	1.38%	0.88%
$0.01\%^2$	(1)	(73)	(3581)	(76917)	(399546)	(684518)	(821589)	(898152)	(946228)
Pitches	66.13%	62.00%	55.88%	40.25%	17.38%	6.50%	3.13%	1.88%	1.38%
$0.1\%^2$	(1)	(73)	(3742)	(96359)	(606175)	(1103560)	(1367417)	(1545154)	(1688526)
Proteins	52.25%	52.13%	51.63%	47.50%	25.25%	4.63%	0.75%	0.25%	0.25%
$0.001\%^2$	(1)	(21)	(422)	(8045)	(128975)	(463357)	(572529)	(589356)	(595906)
Proteins	52.25%	52.13%	51.63%	47.63%	25.75%	5.00%	0.88%	0.50%	0.38%
$0.01\%^2$	(1)	(21)	(422)	(8045)	(131079)	(494846)	(626306)	(654107)	(670114)
Proteins	52.25%	52.13%	51.75%	48.75%	30.13%	7.63%	2.13%	1.50%	1.38%
$0.1\%^2$	(1)	(21)	(426)	(8072)	(143924)	(771311)	(1154106)	(1297080)	(1407901)
Sources	68.75%	47.00%	30.00%	19.75%	14.38%	11.00%	8.50%	6.88%	5.75%
$0.001\%^2$	(1)	(98)	(4557)	(29667)	(75316)	(130527)	(194105)	(259413)	(320468)
Sources	68.75%	47.50%	30.75%	20.13%	14.63%	11.13%	8.63%	7.00%	5.88%
$0.01\%^2$	(1)	(98)	(5615)	(42337)	(103082)	(170646)	(244874)	(320346)	(391369)
Sources	68.75%	51.25%	36.63%	24.38%	16.75%	12.13%	9.13%	7.25%	6.00%
$0.1\%^2$	(1)	(98)	(7372)	(108997)	(319310)	(525914)	(718657)	(904022)	(1080824)

Table 3.9: Empirical entropy statistics for Pseudo-Real Collection (Scheme 2)

3.4.3 Real Texts

Tables 3.10-3.12 give the statistics of artificial texts.

File	Size	Σ	IPM
Cere	440MiB	5	4.301
Para	410MiB	5	4.096
Clostridium Botulium	34MiB	4	3.356
Escherichia Coli	108MiB	15	4.000
Salmonella Enterica	66MiB	9	3.993
Staphylococcus Aureus	38MiB	5	3.579
Streptococcus Pneumoniae	23MiB	8	3.836
Streptococcus Pyogenes	24MIB	10	3.800
Influenza	148MiB	15	3.845
Coreutils	196MiB	236	19.553
Kernel	247MiB	160	23.078
Einstein (en)	446MiB	139	19.501
Einstein (de)	89MiB	117	19.264
Nobel (en)	85MiB	126	20.070
Nobel (de)	31MiB	118	17.786
Turing (en)	7.7MiB	103	21.096
Turing (de)	85MiB	100	19.719
World Leaders	45MiB	89	3.855

Table 3.10: Alphabet statistics for Real Collection

File	p7zip	bzip2	gzip	ppmdi	Re-Pair
Cere	1.14%	2.50%	26.36%	24.09%	1.86%
Para	1.46%	26.34%	27.07%	24.88%	2.80%
Clostridium Botulium	8.53%	25.88%	26.47%	24.12%	20.00%
Escherichia Coli	4.72%	26.85%	28.70%	25.93%	9.63%
Salmonella Enterica	5.61%	27.27%	28.79%	25.76%	12.42%
Staphylococcus Aureus	2.89%	26.32%	28.95%	25.00%	5.26%
Streptococcus Pneumoniae	4.78%	26.52%	27.39%	24.78%	9.57%
Streptococcus Pyogenes	5.00%	26.25%	27.08%	25.00%	9.58%
Influenza	1.35%	6.62%	7.43%	3.78%	3.31%
coreutils	1.94%	16.33%	24.49%	12.76%	2.55%
kernel	0.81%	21.86%	27.13%	18.62%	1.13%
einstein.en	0.07%	5.38%	35.20%	1.61%	0.10%
einstein.de	0.11%	4.38%	31.46%	1.35%	0.16%
nobel.en	0.13%	2.94%	18.82%	1.76%	0.20%
nobel.de	0.18%	3.55%	27.74%	1.68%	0.30%
turing.en	1.09%	36.36%	285.71%	15.58%	1.71%
turing.de	0.03%	0.18%	0.10%	0.11%	0.05%
world leaders	1.29%	7.11%	17.78%	3.56%	1.78%

Table 3.11: Compression statistics for Real Collection

File	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
Cere	27.38% (1)	22.63% (5)	22.63% (25)	22.50% (125)	22.50% (610)	22.50% (2515)	22.50% (8697)	22.38% (28080)	22.25% (88624)
Para	26.50% (1)	23.50% (5)	23.38% (25)	23.38% (125)	23.38% (625)	23.38% (3125)	23.25% (14725)	23.25% (51542)	23.13% (139149)
Clostridium Botulium	23.25% (1)	23.00% (4)	22.88% (16)	22.75% (64)	22.75% (256)	22.75% (1024)	22.63% (4096)	22.50% (16383)	22.25% (65118)
Escherichia Coli	25.00% (1)	24.75% (15)	24.50% (145)	24.38% (779)	24.25% (2715)	24.25% (7436)	24.13% (15641)	24.13% (32561)	23.88% (85363)
Salmonella Enterica	25.00% (1)	24.75% (9)	24.50% (35)	24.38% (97)	24.25% (299)	24.13% (1077)	24.13% (4159)	24.00% (16457)	23.75% (65618)
Staphylococcus Aureus	23.88% (1)	23.75% (5)	23.75% (18)	23.63% (67)	23.63% (260)	23.63% (1029)	23.50% (4102)	23.25% (16391)	22.75% (65282)
Streptococcus Pneumoniae	24.63% (1)	24.38% (8)	24.38% (31)	24.25% (133)	24.13% (574)	24.13% (2183)	24.00% (6928)	23.75% (21093)	23.13% (71592)
Streptococcus Pyogenes	24.50% (1)	24.38% (10)	24.25% (50)	24.13% (174)	24.13% (456)	24.13% (1291)	24.00% (4418)	23.88% (16758)	23.25% (65919)
Influenza	24.63% (1)	24.13% (15)	24.13% (125)	24.00% (583)	23.88% (2329)	23.50% (7978)	22.00% (21316)	18.63% (44748)	13.25% (101559)
coreutils	68.38% (1)	51.25% (236)	35.88% (18500)	23.88% (169716)	17.00% (606527)	12.88% (1335553)	10.13% (2258650)	8.00% (3258896)	6.50% (4247313)
kernel	67.25% (1)	50.50% (160)	36.63% (7122)	25.75% (90396)	19.25% (351918)	15.13% (773818)	12.13% (1305616)	9.63% (1912604)	7.75% (2553008)
einstein.en	62.00% (1)	46.38% (139)	33.38% (4546)	21.13% (28685)	13.25% (77333)	9.00% (142559)	6.50% (211506)	4.75% (276343)	3.50% (335151)
einstein.de	63.00% (1)	44.88% (117)	32.63% (3278)	20.88% (16765)	13.25% (39010)	9.00% (64884)	6.13% (89914)	4.38% (112043)	3.13% (130473)
nobel.en	62.63% (1)	44.63% (126)	30.50% (3566)	18.25% (18079)	11.50% (42334)	8.13% (69855)	6.00% (95644)	4.50% (119260)	3.38% (140401)
nobel.de	61.13% (1)	43.25% (118)	31.13% (2726)	19.63% (12959)	12.50% (30756)	8.63% (49695)	6.00% (66108)	4.13% (80467)	3.00% (92184)
turing.en	63.25% (1)	45.75% (103)	32.00% (2794)	19.13% (14091)	11.50% (33498)	7.63% (55489)	5.38% (75611)	3.88% (93402)	2.88% (108636)
turing.de	62.38% (1)	43.25% (100)	29.25% (1806)	16.75% (7268)	9.50% (15407)	6.00% (23070)	3.88% (29038)	2.63% (33714)	2.00% (37335)
world leaders	43.38% (1)	24.38% (89)	17.25% (2526)	11.63% (23924)	7.63% (106573)	5.13% (246566)	4.00% (374668)	3.50% (468701)	3.13% (547040)

Table 3.12: Empirical entropy statistics for Real Collection