Appendix A: random data input

A.1 Generation of input files

Each input file name has the same format, *e.g.*, the default input file:
`04-08000000-0205-2-04002399-08088848.txt`

The first four numbers are input parameters for the generation of the input file:

- `04`: vector length (number of elements in each vector)
- `08000000`: number of vectors (including duplicates)
- `0205` (`0000` for vector lengths 1 and 2): number of possible values for each vector element*; parameter to adjust internal duplication, which affects the amount of compression possible** (the higher this parameter, the lower the amount of internal duplication)
- `2`: the reciprocal of this number is, approximately, the ratio of duplicate vectors; parameter to adjust the duplication of vectors*** (the higher this parameter, the lower the ratio/number of duplicate vectors)

* each vector element is generated as follows (vector lengths ≥ 3): a number 0..(parameter-1) is randomly generated (possibility of each number is 1/parameter), then this number is "lifted" to an evenly distributed location in a 256MiB hash table of 32-bit integers (to be more precise, the number is multiplied by the factor (4,194,304 / parameter) [evenly distributed selection of bucket (N = 4,194,304) in the 256MiB table, taking 32 as bucket size] * 32 [→ getting start location of bucket (*i.e.*, location of first slot/entry of bucket) in table]; the resulting number is then getting an evenly distributed offset in the selected bucket by this operation: (<resulting number> / 32) % 32) – all /'s are integer divisions
** for different vector lengths and/or number of vectors and/or different ratios of vector duplication, different values for this parameter are needed to get the same amount of compression possible
*** duplicate vectors are randomly distributed throughout the input file

The last two numbers are (implicit) output values from the generation:

- `04002399`: number of unique vectors generated*
- `08088848`: total number of (non-empty) elements in a compressed hash table**

* this number times vector length equals the number of (non-empty) elements in an uncompressed hash table
** this number may be different for different ways of (tree) compression (*e.g.*, regarding the order of elements/nodes in the (implicit) compression tree), for vectors length ≥3;
this number may vary from run to run for vector length 3, due to different hash constants
→ compression ratio = total number of elements in compressed hash table /
total number of elements in uncompressed hash table

A.2 Overview of input files

input file	no. of vectors	dupl.	no. of unique vectors	no. of elements (uncompr.)	no. of elements (compr.)*	compr. ratio
vector length: 1						
01-08000000-0000-2-04000067-08000134.txt						
»	8,000,000	2.00	4,000,067	4,000,067	8,000,134	2
01-08000000-0000-9-07112491-14224982.txt						
»	8,000,000	1.12	7,112,491	7,112,491	14,224,982	2
01-16000000-0000-2-07999225-15998450.txt						
»	16,000,000	2.00	7,999,225	7,999,225	15,998,450	2
01-16000000-0000-9-14224527-28449054.txt						
»	16,000,000	1.12	14,224,527	14,224,527	28,449,054	2
01-32000000-0000-2-15998859-31997718.txt						
»	32,000,000	2.00	15,998,859	15,998,859	31,997,718	2
01-32000000-0000-9-28445271-56890542.txt						
»	32,000,000	1.12	28,445,271	28,445,271	56,890,542	2
vector length: 2						
02-08000000-0000-2-04000852-08001704.txt						
»	8,000,000	2.00	4,000,852	8,001,704	8,001,704	1
02-08000000-0000-9-07111243-14222486.txt						
»	8,000,000	1.12	7,111,243	14,222,486	14,222,486	1
02-16000000-0000-2-08001261-16002522.txt						
»	16,000,000	2.00	8,001,261	16,002,522	16,002,522	1
02-16000000-0000-9-14223349-28446698.txt						
»	16,000,000	1.12	14,223,349	28,446,698	28,446,698	1
vector length: 3						
03-08000000-0600-2-04001692-08723372.txt						
»	8,000,000	2.00	4,001,692	12,005,076	8,723,372	0.73
03-08000000-0825-9-07112578-15586266.txt						
»	8,000,000	1.12	7,112,578	21,337,734	15,586,266	0.73
03-10666666-0700-2-05331050-11642026.txt						
»	10,666,666	2.00	5,331,050	15,993,150	11,642,026	0.73
03-10666666-0925-9-09480500-20672122.txt						
»	10,666,666	1.13	9,480,500	28,441,500	20,672,122	0.73

* this number may be different for different ways of tree compression, for vectors length ≥3; this number may vary from run to run for vector length 3, due to different hash constants

table continues on next page…

vector length: 4						
04-04000000-0200-2-01999584-04079168.txt						
»	4,000,000	2.00	1,999,584	7,998,336	4,079,168	0.51
04-04000000-0200-9-03555969-07191938.txt						
»	4,000,000	1.13	3,555,969	14,223,876	7,191,938	0.51
04-04000000-0965-2-02000436-05837892.txt						
»	4,000,000	2.00	2,000,436	8,001,744	5,837,892	0.73
04-04000000-1290-9-03554818-10391626.txt						
»	4,000,000	1.12	3,554,818	14,219,272	10,391,626	0.73
04-08000000-0205-2-04002399-08088848.txt (default)						
»	8,000,000	2.00	4,002,399	16,009,596	8,088,848	0.51
04-08000000-0275-9-07110687-14372624.txt						
»	8,000,000	1.13	7,110,687	28,442,748	14,372,624	0.51
04-08000000-1365-2-03999728-11675190.txt						
»	8,000,000	2.00	3,999,728	15,998,912	11,675,190	0.73
04-08000000-1820-9-07111327-20756906.txt						
»	8,000,000	1.12	7,111,327	28,445,308	20,756,906	0.73
04-16000000-0300-2-08000672-16181344.txt						
»	16,000,000	2.00	8,000,672	32,002,688	16,181,344	0.51
04-16000000-0400-9-14220385-28760770.txt						
»	16,000,000	1.13	14,220,385	56,881,540	28,760,770	0.51
04-16000000-1920-2-08004559-23286302.txt						
»	16,000,000	2.00	8,004,559	32,018,236	23,286,302	0.73
04-16000000-2570-9-14221639-41474428.txt						
»	16,000,000	1.13	14,221,639	56,886,556	41,474,428	0.73

table continues on next page…

vector length: 8						
08-04000000-0040-2-02001551-08053058.txt						
»	4,000,000	2.00	2,001,551	16,012,408	8,053,058	0.50
08-04000000-0047-9-03554816-14602198.txt						
»	4,000,000	1.13	3,554,816	28,438,528	14,602,198	0.51
08-04000000-0080-2-02000449-11637432.txt						
»	4,000,000	2.00	2,000,449	16,003,592	11,637,432	0.73
08-04000000-0093-9-03555564-20695988.txt						
»	4,000,000	1.12	3,555,564	28,444,512	20,695,988	0.73
08-08000000-0048-2-04000799-16264610.txt						
»	8,000,000	2.00	4,000,799	32,006,392	16,264,610	0.51
08-08000000-0055-9-07111271-28660530.txt						
»	8,000,000	1.12	7,111,271	56,890,168	28,660,530	0.50
08-08000000-0094-2-04000463-23227868.txt						
»	8,000,000	2.00	4,000,463	32,003,704	23,227,868	0.73
08-08000000-0110-9-07110432-41347914.txt						
»	8,000,000	1.13	7,110,432	56,883,456	41,347,914	0.73