Profiling

Caden Corontzos

I used callgrind to profile my code. Callgrind finds out how many instructions are associated with each line of code, which gives an indication of how long each line will take. I looked at some of the trouble spots that Eitan and I found to see if we could see a noticable difference in runtime before and after each change.

I used the two groups of DNA I had found the other day as benchmarks.

Here is the data before any optimizations.

Table 1: Corpus 1

File.Name	Original.File.Size	Compressed.Size	Compression.Ratio	Compresssion.Time	Decompression.Time
DNACorpus1/chmpxx	121024	43516	2.781	55	18
DNACorpus1/chntxx	155844	58336	2.671	77	27
DNACorpus1/hehcmv	229354	85526	2.682	118	36
DNACorpus1/humdyst	38770	15300	2.534	35	7
DNACorpus1/humghcs	66495	25552	2.602	33	12
DNACorpus1/humhbb	73308	28134	2.606	37	12
DNACorpus1/humhdab	58864	22699	2.593	30	32
DNACorpus1/humprtb	56737	21902	2.590	32	17
DNACorpus1/mpomtcg	186609	70254	2.656	122	46
DNACorpus1/mtpacga	100314	36862	2.721	73	18
DNACorpus1/vaccg	191737	70067	2.736	97	30

Table 2: Corpus 2

File.Name	Original.File.Size	Compressed.Size	Compression.Ratio	Compresssion.Time	Decompression.Time
DNACorpus2/AeCa	1591049	556535	2.859	1221	289
DNACorpus2/AgPh	43970	17442	2.521	21	8
DNACorpus2/AnCa	142189675	43665091	3.256	160561	33570
DNACorpus2/BuEb	18940	7893	2.400	34	7
DNACorpus2/DaRe	62565020	19586457	3.194	58734	16327
DNACorpus2/DrMe	32181429	10619042	3.031	29080	9883
DNACorpus2/EnIn	26403087	8609993	3.067	23231	7525
DNACorpus2/EsCo	4641652	1593404	2.913	3372	1110
DNACorpus2/GaGa	148532294	46851765	3.170	141631	27250
DNACorpus2/HaHi	3890005	1306708	2.977	3067	955
DNACorpus2/HePy	1667825	566972	2.942	1251	271
DNACorpus2/HoSa	189752667	57200209	3.317	168619	35951
DNACorpus2/OrSa	43262523	14148071	3.058	31413	7275
DNACorpus2/PlFa	8986712	2895744	3.103	6721	2041
DNACorpus2/ScPo	10652155	3590856	2.966	8229	2277
DNACorpus2/WaMe	9144432	3112000	2.938	6779	1994
DNACorpus2/YeMi	73689	27235	2.706	38	18

I then tested my implementation with callgrind. I encoded HaHi from DNA Corpus 2 to see what lines are taking long.

The first change we want to make was to create a type for codewords. They were previously just int, but we want to make them fixed at uint64 and have a special type for them.

I ran callgrind before and after making this change. I found that we were able to save a lot of instructions on dictionary accesses with the fixed type. For instance, this was the number of

```
{c++, eval= F}   6,910,075 ( 0.01%)              dictionary[currentBlock + next_character]
= codeword;
```

Table 3: Corpus 1

File.Name	Original.File.Size	Compressed.Size	Compression.Ratio	Compresssion.Time	Decompression.Time
DNACorpus1/chmpxx	121024	43516	2.781	60	22
DNACorpus1/chntxx	155844	58336	2.671	77	30
DNACorpus1/hehcmv	229354	85526	2.682	120	50
DNACorpus1/humdyst	38770	15300	2.534	20	16
DNACorpus1/humghcs	66495	25552	2.602	33	15
DNACorpus1/humhbb	73308	28134	2.606	35	23
DNACorpus1/humhdab	58864	22699	2.593	34	14
DNACorpus1/humprtb	56737	21902	2.590	32	19
DNACorpus1/mpomtcg	186609	70254	2.656	114	30
DNACorpus1/mtpacga	100314	36862	2.721	54	21
DNACorpus1/vaccg	191737	70067	2.736	96	28

Table 4: Corpus 2

File.Name	Original.File.Size	Compressed.Size	Compression.Ratio	Compresssion.Time	Decompression.Time
DNACorpus2/AeCa	1591049	556535	2.859	1027	359
DNACorpus2/AgPh	43970	17442	2.521	27	10
DNACorpus2/AnCa	142189675	43665091	3.256	130129	30638
DNACorpus2/BuEb	18940	7893	2.400	11	7
DNACorpus2/DaRe	62565020	19586457	3.194	55746	14370
DNACorpus2/DrMe	32181429	10619042	3.031	28462	5926
DNACorpus2/EnIn	26403087	8609993	3.067	22359	7255
DNACorpus2/EsCo	4641652	1593404	2.913	3806	1415
DNACorpus2/GaGa	148532294	46851765	3.170	141216	31454
DNACorpus2/HaHi	3890005	1306708	2.977	3001	692
DNACorpus2/HePy	1667825	566972	2.942	1197	333
DNACorpus2/HoSa	189752667	57200209	3.317	182781	41298
DNACorpus2/OrSa	43262523	14148071	3.058	35052	7914
DNACorpus2/PlFa	8986712	2895744	3.103	5889	2225
DNACorpus2/ScPo	10652155	3590856	2.966	8702	2699
DNACorpus2/WaMe	9144432	3112000	2.938	6045	1814
DNACorpus2/YeMi	73689	27235	2.706	36	14