



hochschule mannheim

Comparison of compression tools for biological data and analysis of possible improvements

Gabriel Eichelkraut

Bachelor Thesis

for the acquisition of the academic degree Bachelor of Science (B.Sc.)

Course of Studies: Computer Science

Department of Computer Science

University of Applied Sciences Mannheim

01.12.22

Tutors

Prof. Elena Fimmel, Hochschule Mannheim

TBD

Eichelkraut, Gabriel:

Comparison of compression tools for biological data and analysis of possible improvements

/ Gabriel Eichelkraut. –

Bachelor Thesis, Mannheim: University of Applied Sciences Mannheim, 2022. 7 pages.

Eichelkraut, Gabriel:

Vergleich von Kompressionswerkzeugen für biologische Daten und Analyse von Verbesserungsmöglichkeiten

/ Gabriel Eichelkraut. –

Bachelor-Thesis, Mannheim: Hochschule Mannheim, 2022. 7 Seiten.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

This work is licensed under a Creative Commons
“Attribution-ShareAlike 4.0 International” license.



Mannheim, 01.12.22

Gabriel Eichelkraut

Abstract

Comparison of compression tools for biological data and analysis of possible improvements

The European languages are members of the same family. Their separate existence is a myth. For science, music, sport, etc, Europe uses the same vocabulary. The languages only differ in their grammar, their pronunciation and their most common words. Everyone realizes why a new common language would be desirable: one could refuse to pay expensive translators. To achieve this, it would be necessary to have uniform grammar, pronunciation and more common words. If several languages coalesce, the grammar of the resulting language is more simple and regular than that of the individual languages. The new common language will be more simple and regular than the existing European languages. It will be as simple as Occidental; in fact, it will be Occidental. To an English person, it will seem like simplified English, as a skeptical Cambridge friend of mine told me what Occidental is.

Vergleich von Kompressionswerkzeugen für biologische Daten und Analyse von Verbesserungsmöglichkeiten

Jemand musste Josef K. verleumdet haben, denn ohne dass er etwas Böses getan hätte, wurde er eines Morgens verhaftet. Wie ein Hund! sagte er, es war, als sollte die Scham ihn überleben. Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. Und es war ihnen wie eine Bestätigung ihrer neuen Träume und guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. Es ist ein eigentümlicher Apparat, sagte der Offizier zu dem Forschungsreisenden und überblickte mit einem gewissermaßen bewundernden Blick den ihm doch wohl bekannten Apparat. Sie hätten noch ins Boot springen können, aber der Reisende hob ein schweres, geknotetes Tau vom Boden, drohte ihnen damit und hielt sie dadurch von dem Sprunge ab. In den letzten Jahrzehnten ist das Interesse an Künstlern sehr zurückgegangen. Aber sie überwandten sich, umdrängten den Käfig und wollten sich gar nicht fortrühren.

Contents

1. Introduction	1
2. Structure Of Biological Data	3
2.1. DNA	3
2.2. Nucleotides	3
3. Datatypes	5
3.0.1. FASTQ	5
3.0.2. SAM/BAM	6
3.1. Compression aproaches	7
3.1.1. Huffman encoding	7
List of Abbreviations	vii
List of Tables	ix
List of Figures	xi
Listings	xiii
Index	xvii
A. Erster Anhang: Lange Tabelle	xvii

Chapter 1

Introduction

Chapter 2

Structure Of Biological Data

To strengthen the understanding how and where biological information is stored, this section starts with a quick and general rundown on the structure of any living organism. All living organisms, like plants and animals, are made of cells (a human body can consist out of several trillion cells). A cell in itself is a living organism, the smallest one possible. A cell has two layers, the inner one is called nucleus which contains chromosomes. The chromosomes hold the genetic information in form of DNA.

2.1. DNA

Deoxyribonucleic acid (DNA) is often seen in the form of a double helix. A double helix consists, as the name suggests, of two single helix. Each of them consists of two main components: the Sugar Phosphate backbone, which is irrelevant for this Paper and the Bases. The arrangement of Bases represents the Information stored in the DNA. A base is an organic molecule, they are called Nucleotides.

2.2. Nucleotides

For this paper, nucleotides are the most important parts of the DNA. A Nucleotide can have one of four forms: it can be either adenine, thymine, guanine or cytosine. Each of them has a Counterpart on the helix, to be more explicit: adenine can only bond with thymine, guanine can only bond with cytosine. This means with

2. Structure Of Biological Data

the content of one helix, the other one can be determined by “inverting” the first. The counterpart for e.g.: adenine, guanine, adenine would be: thymine, cytosine, thymine. For the sake of simplicity, one does not write out the full name of each nucleotide but only use its initial: AGA in one Helix, TCT in the other.

Chapter 3

Datatypes

DNA can be represented by a String with the buildingblocks A,T,G and C. Using a common fileformat for saving text would be impractical because the encoding defines that other possible letters require more space per letter. So storing a single A in ASCII encoding requires 4 bit (excluding the magic bytes in the fileheader), whereas only two bits are needed to save a letter with a four letter alphabet e.g.: 00 -> A, 01 -> T, 10 -> G, 11 -> C. More common Text encodings like unicode require even more storage space per letter. So settling with ASCII has improvement capabilities but is, on the other side, more efficient than using bulkier alternatives.

To optimize this, people have developed other filetypes, that focus on storing nucleotides only. Standard formats include:

- FASTQ
- SAM/BAM

Since methods to store this kind of Data are still in development, there are many more filetypes. The few, mentioned above are used by different organisations and researchers and backed by a scientific publication. In order to not go beyond the scope, this paper will only focus on compression tools which are using standard formats.

3.0.1. FASTQ

Is a text based format for storing sequenced data. It saves nucleotides as letters and in extend to that, the quality values are saved. FASTQ files are split into multiples

3. Datatypes

of four, each four lines contain the informations for one sequence. The exact structure of FASTQ format is as follows: Line 1: Sequence identifier aka. Title, starting with an @ and an optional description. Line 2: The sequence consisting of nucleoids, symbolized by A, T, G and C. Line 3: A '+' that functions as a separator. Optionally followed by content of Line 1. Line 4: quality line(s). consisting of letters and special characters in the ASCII scope.

The quality values have no fixed type, to name a few there is the sanger format, the solexa format introduced by Solexa Inc., the Illumina and the QUAL format, generated by the PHRED software. The quality value shows the estimated probability of error in the sequencing process.

3.0.2. SAM/BAM

SAM Sequence Alignment/Map format, often just called BAM like its fileextension, is part of the SAMtools package, a utility tool for processing SAM/BAM and CRAM files. The SAM/BAM file is a text based format delimited by TABs. It uses 7-bit US-ASCII, to be precise Charset ANSI X3.4-1968 as defined in RFC1345. The structure is more complex than the one in FASTQ and described best, accompanied by an example:

```
Coor      12345678901234 5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002      aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004      ATAGCT.....TCAGC
-r003      ttagctTAGGC
-r001/2      CAGCGGCAT
```

Figure 3.1.: SAM/BAM file structure example

Valid Symbols

The regular expression, shown above, filters tuple of characters from a to z in lower and uppercase, numbers and several special characters. There are minor dif-

ferences between the special chars first and second element of the tuple, in the latter '*' and '=' are allowed which are not allowed in the first.

Index File

- may contain metadata: magic string, ref seq, bins, offsets, unmapped reads - allows viewing BAM data (locally and remote via ftp/http) - file extension: <filename>.bam.bai
- saves more data than FASTQ
- alignment section includes - RG Read Group - BC Barcode Tag - SM Single-end alignment quality - AS Paired-end alignment quality - NM Edit distance tag - XN amplicon name tag
- BAM index files nameschema: <filename>.bam.bai

3.1. Compression approaches

The process of compressing data serves the goal to generate an output that is smaller than its input data. In many cases, like in gene compressing, the compression is ideally lossless. This means it is possible for every compressed data, to receive the full information that were available in the origin data, by decompressing it. Lossy compression on the other hand, might excludes parts of data in the compression process, in order to increase the compression rate. The excluded parts are typically not necessary to transmit the origin information. This works with certain audio and pictures files or network protocols that are used to transmit video/audio streams live. For DNA a lossless compression is needed. To be precise a lossy compression is not possible, because there is no unnecessary data. Every nucleotide and its position is needed for the sequenced DNA to be complete.

The well known Huffman coding, is used in several Tools for genome compression (genomic squeeze <- official | unofficial -> GDC, GRS). Further Arithmetic numeral system (ANS) or rANS ... TBD.

3.1.1. Huffman encoding

List of Abbreviations

DNA Deoxyribonucleic acid
ANS Arithmetic numeral system

List of Tables

A.1. Tabelle mit ISO-Ländercodes	xvii
--	------

List of Figures

3.1. SAM/BAM file structure example	6
---	---

Listings

Appendix A

Erster Anhang: Lange Tabelle

Hier ein Beispiel für einen Anhang. Der Anhang kann genauso in Kapitel und Unterkapitel unterteilt werden, wie die anderen Teile der Arbeit auch.

Table A.1.: Lange Tabelle mit ISO-Ländercodes

Country	A 2	A 3	Number
AFGHANISTAN	AF	AFG	004
ALBANIA	AL	ALB	008
ALGERIA	DZ	DZA	012
AMERICAN SAMOA	AS	ASM	016
ANDORRA	AD	AND	020
ANGOLA	AO	AGO	024
ANGUILLA	AI	AIA	660
ANTARCTICA	AQ	ATA	010
ANTIGUA AND BARBUDA	AG	ATG	028
ARGENTINA	AR	ARG	032
ARMENIA	AM	ARM	051
ARUBA	AW	ABW	533
AUSTRALIA	AU	AUS	036
AUSTRIA	AT	AUT	040
AZERBAIJAN	AZ	AZE	031
BAHAMAS	BS	BHS	044
BAHRAIN	BH	BHR	048
BANGLADESH	BD	BGD	050
BARBADOS	BB	BRB	052
BELARUS	BY	BLR	112
BELGIUM	BE	BEL	056
BELIZE	BZ	BLZ	084
BENIN	BJ	BEN	204
BERMUDA	BM	BMU	060
BHUTAN	BT	BTN	064
BOLIVIA	BO	BOL	068
BOSNIA AND HERZEGOWINA	BA	BIH	070
BOTSWANA	BW	BWA	072
BOUVET ISLAND	BV	BVT	074
BRAZIL	BR	BRA	076
BRITISH INDIAN OCEAN TERRITORY	IO	IOT	086
BRUNEI DARUSSALAM	BN	BRN	096
BULGARIA	BG	BGR	100
BURKINA FASO	BF	BFA	854
BURUNDI	BI	BDI	108
CAMBODIA	KH	KHM	116
CAMEROON	CM	CMR	120

A. Erster Anhang: Lange Tabelle

CANADA	CA	CAN	124
CAPE VERDE	CV	CPV	132
CAYMAN ISLANDS	KY	CYM	136
CENTRAL AFRICAN REPUBLIC	CF	CAF	140
CHAD	TD	TCD	148
CHILE	CL	CHL	152
CHINA	CN	CHN	156
CHRISTMAS ISLAND	CX	CXR	162
COCOS (KEELING) ISLANDS	CC	CCK	166
COLOMBIA	CO	COL	170
COMOROS	KM	COM	174
CONGO	CG	COG	178
COOK ISLANDS	CK	COK	184
COSTA RICA	CR	CRI	188
COTE D'IVOIRE	CI	CIV	384
CROATIA (local name: Hrvatska)	HR	HRV	191
CUBA	CU	CUB	192
CYPRUS	CY	CYP	196
CZECH REPUBLIC	CZ	CZE	203
DENMARK	DK	DNK	208
DJIBOUTI	DJ	DJI	262
DOMINICA	DM	DMA	212
DOMINICAN REPUBLIC	DO	DOM	214
EAST TIMOR	TP	TMP	626
ECUADOR	EC	ECU	218
EGYPT	EG	EGY	818
EL SALVADOR	SV	SLV	222
EQUATORIAL GUINEA	GQ	GNQ	226
ERITREA	ER	ERI	232
ESTONIA	EE	EST	233
ETHIOPIA	ET	ETH	210
FALKLAND ISLANDS (MALVINAS)	FK	FLK	238
FAROE ISLANDS	FO	FRO	234
FIJI	FJ	FJI	242

Beachten Sie, dass die Tabelle manchmal erst nach dreimaligem Lauf durch $\LaTeX$ richtig angezeigt wird.