Sequence analysis

Cryfa: a secure encryption tool for genomic data

Morteza Hosseini*, Diogo Pratas and Armando J. Pinho

IEETA, Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810–193 Aveiro, Portugal

*To whom correspondence should be addressed.

Abstract

Summary: The ever-increasing growth of high-throughput sequencing technologies has led to a great acceleration of medical and biological research and discovery. As these platforms advance, the amount of information for diverse genomes increases at unprecedented rates. Confidentiality, integrity and authenticity of such genomic information should be ensured due to its extremely sensitive nature. In this paper, we propose Cryfa, a fast secure encryption tool for genomic data, namely in Fasta, Fastq, VCF, SAM and BAM formats, which is also capable of reducing the storage size of Fasta and Fastq files. Cryfa uses advanced encryption standard (AES) encryption combined with a shuffling mechanism, which leads to a substantial enhancement of the security against low data complexity attacks. Compared to AES Crypt, a general-purpose encryption tool, Cryfa is an industry-oriented tool, which is able to provide confidentiality, integrity and authenticity of data at four times more speed; in addition, it can reduce the file sizes to 1/3. Due to the absence of a method similar to Cryfa, we have simulated its behavior with a combination of encryption and compression tools, for comparison purpose. For instance, our tool is nine times faster than its fastest competitor in Fasta files. Also, Cryfa has a very low memory usage (only a few megabytes), which makes it feasible to run on any computer.

Availability and implementation: Source codes and binaries are available, under GPLv3, at https://github.com/pratas/cryfa.

Contact: seyedmorteza@ua.pt

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The rapid advancement in high-throughput sequencing technologies has triggered a revolution in personalized medicine, biotechnology and ancient DNA studies (Kumar-Sinha and Chinnaiyan, 2018; Porter and Hajibabaei, 2018). However, it raises critical issues regarding preserving security of the genomic data, which is highly sensitive due to its nature.

Authenticated encryption has the potential to address the issues of genomic data security, by allowing only authorized parties to access such data. This cryptographic scheme is able to appropriately ensure the key criteria of integrity, confidentiality and authenticity of the genomic data, under a single and easy-to-use programing interface (Bradley et al., 2017; Jagadeesh et al., 2017). Secure encryption of genomic data is of great significance not only for macro-organisms, e.g. humans, but also for microbial species, such as outbreak microorganisms, e.g. Ebola and Zika viruses (Graham and Sullivan, 2018), ancient viruses (Duggan et al., 2016), exobiology sources (Martins et al., 2008), forensics applications (Budowle et al., 2014) and synthetic biology (Deane, 2018).

General-purpose encryption methods, although directly applicable, do not take into consideration specific properties of genomic data files; for instance, Fasta files contain headers, beginning with the ‘>’ character, and DNA bases, including ‘A’, ‘C’, ‘G’, ‘T’ and ‘N’ symbols; Fastq files comprise headers, beginning with the ‘@’ character, bases, ‘+’ separators and quality scores; VCF files begin with ‘##fileformat=VCF’ string. Thus, a specific-purpose encryption approach is required, that is able to protect the genomic data against known-plaintext attacks (KPA) (Zhang et al., 2018), as well as low data complexity attacks (Bouillaguet et al., 2012).
In this paper, we present the Cryfa tool, that follows industry recommendations for upholding security of in-transit and at-rest genomic data. This tool addresses secure encryption of such data, along with compacting Fasta/Fastq sequences by a block transformation, followed by shuffling the transformed information and ultimately, performing a fast authenticated encryption on the shuffled content. The encryption is performed with the advanced encryption standard (AES), which is announced by the US. National Institute of Standards and Technology. AES is a symmetric-key algorithm; it uses cipher keys to process data blocks based on a substitution-permutation network (Daemen and Rijmen, 2002).

Applying the AES method distributes the information uniformly, thus it is needed to perform the compacting phase before encryption. We perform a fixed-size compacting, i.e. we pack equally sized blocks of symbols in Fasta/Fastq files, independently of their redundancy. Moreover, operating the shuffling before encryption is crucial, since it prevents an adversary to break the encryption by low data complexity or KPA attacks. Also, it enormously increases the time for breaking encryption by performing an exhaustive search on the password. For a discussion on the importance of applying the shuffling phase, see Supplementary Note S1.2.

2 Materials and methods

The schema of Cryfa is demonstrated in Supplementary Figure S1. For the purpose of encrypting and compacting a Fastq file by Cryfa, it is first split into headers, bases and quality scores. Similarly, a Fasta file is split into headers and bases. In the next step, packing of these split segments is performed in different fixed-size blocks, in a way that each block maps a tuple of symbols into an ASCII character. The number of symbols considered for each tuple can be different for headers, bases and quality scores. The next step is employing a key file, containing a password, to shuffle the packed content that is obtained by joining the outputs of different packing blocks. Supplementary Note S4 provides with a guideline for making the key file, which can be carried out by the ‘keygen’ tool that we have provided alongside the Cryfa tool. As the result of shuffling, the content becomes uniformly permuted and transformed into pseudo high-data complexity; hence, it becomes resistant against low data complexity and KPA attacks. In the final step, an authenticated encryption, which simultaneously provides data confidentiality and integrity, is carried out on the shuffled content, by the AES method in Galois/counter mode (GCM). The output of this final step is an encrypted and compact Fasta/Fastq file.

In order to decrypt and unpack a file, it is first decrypted by the AES method in GCM mode. Then, the decrypted content is unshuffled using the key file that is restored to order from the shuffled state. Note that the key file used in this phase needs to be the same as the one used for shuffling. Finally, the unshuffled content is unpacked using a lookup table, and the decrypted and unpacked file is obtained. This file is the same as the original Fasta/Fastq file which had been encrypted and compacted, due to the lossless nature of the Cryfa tool.

Cryfa is capable of preserving the privacy of any genomic data in Fasta, Fastq, VCF, SAM and BAM formats. In this way, if a genomic file, e.g. in VCF format, is passed to Cryfa, it can be efficiently shuffled and encrypted. Supplementary Note S1 describes the methods in greater detail.

3 Results

We have tested Cryfa and AES Crypt, a general-purpose encryption software, on a collection of Fasta, Fastq, VCF, SAM and BAM datasets. A description of the datasets used, the results obtained and a manual for running Cryfa are provided in Supplementary Notes S2–S4, respectively. The results of comparing Cryfa and AES Crypt are demonstrated in Supplementary Tables S1 and S2 and Supplementary Figure S7. Although Cryfa shuffles, efficiently preserves security and integrity and compact the files, it is 2.2–4 times faster than AES Crypt on Fasta/Fastq datasets. Moreover, it is able to reduce 47–66% of the space required to store a file, whilst AES Crypt does not compress the data. Also, Cryfa performs shuffling and secure encryption of VCF/SAM/BAM datasets at 1.7–2.6 times more speed than AES Crypt. In this case, Cryfa uses only 1 MB of RAM.

We have carried out several general-purpose and specific-purpose compression plus encryption on the mentioned Fasta and Fastq datasets. Comparing our tool with these methods, demonstrated in Supplementary Tables S6–S8 and Supplementary Figure S8, Cryfa is nine times faster than its fastest competitor on Fasta datasets that is bzip2 plus AES Crypt. This value is 1.4 for Fastq datasets, compared with DSRC 2 plus AES Crypt, due to the more complex nature of the Fastq format. Note that since there was no method that can simultaneously compress and encrypt genomic data, similar to what Cryfa does, we simulated Cryfa’s behavior with a combination of compression methods and an encryption tool.

To evaluate the cost-effectiveness of running Cryfa on multi-core computing resources, e.g. in the cloud, we have carried it out using a different number of threads on two sample datasets. The results are shown in Supplementary Figure S9. Running with eight threads, compared to one thread, is 2.4 times faster, in terms of real time, and it takes 1.4 times more CPU time, as an aggregation of user and system times, on average. The difference between memory usages while running with one thread and eight threads is 10 MB, which is insignificant. Cryfa uses, at most, 31 MB of RAM.

State-of-the-art genomic sequence compressors explore redundancy in files to further compress them. This can be exploited by security attacks to differentiate species. Supplementary Figure S10 shows that DELIMINATE and MFCompress have diverse normalized compression values, which makes them vulnerable against the mentioned exploitations. Contrarily, Cryfa behaves similarly in datasets from different species, since it does not explore redundancy in the files. This way, Cryfa is able to further preserve the confidentiality of genomic data. For a more detailed description of results, see Supplementary Note S3.

In conclusion, we have proposed Cryfa, an industry-oriented tool to securely encrypt genomic data in Fasta/Fastq/VCF/SAM/BAM formats, and also, compact data in Fasta/Fastq formats. The security of such data is substantially improved by a straightforward mechanism, shuffling. We further preserve the security of genomic data by not exploring complexity in those files. Therefore, Cryfa cannot be exploited for species differentiation. Our tool is approximately one order of magnitude faster than the fastest state-of-the-art compression plus encryption tools, including general-purpose and specific-purpose ones. Cryfa not only is high-speed and provides a high level of security, as also has a very low memory usage (only a few megabytes).

Funding

This work was supported by European Fund for Regional Development (FEDER) through the Operational Program Competitiveness Factors (COMPETE); and by national funds through the Portuguese Foundation for Science and Technology (FCT), in the context of the projects [UID/CEC/
Conflict of Interest: none declared.

References

Bouillaguet,C. et al. (2012) Low-data complexity attacks on AES. IEEE Trans. Inf. Theory, 58, 7002–7017.

Bradley,T. et al. (2017) Genomic security (lest we forget). IEEE Security Privacy Mag., 15, 38.

Budowle,B. et al. (2014) Validation of high throughput sequencing and microbial forensics applications. Investig. Genet., 5, 9.

Daemen,J. and Rijmen,V. (2002). The Design of Rijndael: AES—the Advanced Encryption Standard. Springer-Verlag, Berlin, Heidelberg.

Deane,C. (2018) Synthetic biology: license to kill. Nat. Chem. Biol., 14, 107.

Duggan,A.T. et al. (2016) 17th century variola virus reveals the recent history of smallpox. Curr. Biol., 26, 3407–3412.

Graham,B.S. and Sullivan,N.J. (2018) Emerging viral diseases from a vaccinology perspective: preparing for the next pandemic. Nature Immunol., 19, 20.

Jagadeesh,K.A. et al. (2017) Deriving genomic diagnoses without revealing patient genomes. Science, 357, 692–695.

Kumar-Sinha,C. and Chinnaiyan,A.M. (2018) Precision oncology in the age of integrative genomics. Nature Biotechnol., 36, 46.

Martins,Z. et al. (2008) Extraterrestrial nucleobases in the murchison meteorite. Earth Planet. Sci. Lett., 270, 130–136.

Porter,T.M. and Hajibabaei,M. (2018) Scaling up: a guide to high throughput genomic approaches for biodiversity analysis. Mol. Ecol., 27, 313–338.

Zhang,L.Y. et al. (2018) Improved known-plaintext attack to permutation-only multimedia ciphers. Inf. Sci., 430–431, 228–239.