GENOME_COMPRESSOR : Un compresseur bio-inspiré pour les données textuelles, génomiques et cybernétique

GENOME_COMPRESSOR : A Bio-Inspired Compressor for Textual, Genomic, and Cyber Data

Rakotondravelo Tahina Mickaël

Étudiant autodidacte - Centre National de Télé-Enseignement de Madagascar (CNTEMAD)

Darkphex@gmail.com

Résumé

GENOME_COMPRESSOR est un projet open source proposant un compresseur de données inspiré du vivant. Il repose sur l'encodage ADN pour structurer et compresser des données complexes, qu'il s'agisse de génomes, de textes ou de journaux informatiques. L'approche repose sur une modularisation fine (détection de motifs, encodage génétique, gestion des mutations) et un format portable .dna. Le projet vise une double finalité : fournir un outil de compression performant, mais également offrir une base pédagogique et expérimentale pour explorer les liens entre bioinformatique, compression et cybersécurité.

Abstract

GENOME_COMPRESSOR is an open-source project offering a data compressor inspired by biological systems. It relies on DNA encoding to structure and compress complex data, including genomes, texts, or log files. the approach is based on fine modularization (pattern detection, genetic encoding, mutation handling) and a portable .dna format. The project has a dual purpose: providing a high-performance compression tool and offering an educational and experimental base to explore the links between bioinformatics, compression, and cybersecurity.

Mots-clés : compression, ADN, bio-inspiration, génomique, cybersécurité, open source, journaux système, données textuelles

Keywords : compression, DNA, bio-inspiration, genomics, cybersecurity, open source, system logs, textual data

1 Introduction

Au cours des dernières décennies, le volume de données numériques génerées à l'échelle mondiale a connu une croissance exponentielle, touchant tous les secteurs, du médical à l'industrie, en passant par les communications, la recherche et la cybersécurité. Face à cette explosion, deux enjeux majeurs émergent : la nécessité de compresser efficacement ces volumes toujours croissants et celle de garantir la sécurité des informations sensibles qu'ils contiennent.

Dans ce contexte, les solutions classiques de compression (telles que gzip, bizp2 ou 7zip) montrent des limites, notamment lorsqu'il s'agit de données non structurées ou buitées, ou encore dans le contextes où la traçabilité, l'intégrité ou le chiffrement sont requis.

Le projet GENOME_COMPRESSOR s'inscrit dans une démarche innovante mêlant bio-inspiration et cybersécurité légère. Il explore les analogies entre l'encodage et l'information génétique et les défis du traitement de données modernes, afin de proposer une alternative open source à la fois originale, modulaire et extensible.

Ce compresseur bio-inspiré est actuellement publié en licence libre MIT. Cet article présente la conception, les fondements théoriques, et les performances actuelles de la version open source de `GENOME_COMPRESSOR`, tout en exposant les perspectives futures en matière d'optimisation et d'enrichissement fonctionnel.

2 Concepts et Inspirations

Le projet GENOME_COMPRESSOR s'appuis sur une approche bio-inspirée, à la croisée des sciences du vivant et de l'informatique. L'idée centrale repose sur l'observation que l'ADN, support de l'information génétique chez les être vivants, constitue un système de stockage et de transmission de données extrêmement compact, résilient et universel. Cette observation a motivé une réflexion sur l'adaptabilité de ces principes au domaine de la compression de données.

2.1 Compression bio-inspirée

La compression bio-inspirée s'inscrit dans le champ plus large de l'algorithme naturelle. ELle vise à imiter les mécanismes d'organisation de répétition et de variation présents dans les séquences génétiques pour identifier et exploiter les régularités dans les données numeriques. Cette approche permet de capturer des motifs, des structures redondantes ou auto-similaires que les algorithme classiques détectent parfois difficilement, en particulier dans les jeux de données non structures ou bruitées.

2.2 Encodage ADN

L'encodage ADN, tel qu'implémenté dans GENOME_COMPRESSOR, repose sur la transposition symbolique des données vers un alphabet à quatre lettre, inspiré des bases nucléiques (A, C, G, T). Ce principe permet une représentation stable, linéaire et compatible avec des manipulations bio-informatiques. L'encodage ADN présente également des avantages en matière de resilience aux erreurs, de traçabilité et de modularité, tout en conservant une faible empreinte mémoire.

2.3 Inspirations en cybersécurité

La cybersécurité constitue un pilier fondamental de l'evolution du projet. Plutôt que de recourir uniquement à des techniques lourdes de chiffrement GENOME_COMPRESSOR s'inspire de principes biologiques pour favoriser la discrétion, la segmentation de l'information et la transformation non triviale des données. Cette stratégie vise à retarder l'identification des structures sensibles, à faciliter l'intégration de modules de chiffrement légers et à conserver une grande portabilité sur differents systèmes.

3 Architecture du système (version actuelle)

La version actuelle de GENOME_COMPRESSOR est conçue selon une architecture modulaire, chaque composant remplissant une fonction spécifique dans le pipeline

de compression et de décompression. Cette organisation vise à assurer une clarté fonctionnelle, une facilité de maintenace et une extensibilité pour les futures versions.

3.1 Modules principaux

Les système repose sur une série de modules spécialisés, interopérables, codés en python :

- `pattern_scanner.py` : détecte les motifs récurrents dans les données d'entrée en s'appuyant sur des algorithmes optimisés de type Rabin-Karp
- `gene_encoder.py` : encode les motifs identifiés sous une forme génétique simulée, en les traduisant dans un alphabet inspiré de l'ADN.
- `mutation_encoder.py` : gère les variations locales dans les données (bruit, erreurs, anomalies) en les représentant comme des mutations contrôlées.
- `genome_compressor` : orchestre l'ensemble du processus de compression, en appelant les modules précedents et en structurant les sorties
- `storage_model.py` : assure la sauvegarde des fichiers .dna dans un format JSON structuré, versionné, portable et lisible.
- `genome_decoder.py` : réalise l'opération inverse complète, en restaurant les données originales à partir d'un fichier .dna

3.2 Pipeline de traitement

Le pipeline standard de GENOME_COMPRESSOR suit les étapes suivantes :

1. Analyse des données : extraction de motifs fréquents et évaluation de la structure locale.
2. Encodage génétique : représentation des motifs et de leur variation sous forme de "gènes numériques".
3. Écriture du fichier compressé : génération du fichier .dna, structuré, compact et potentiellement sécurisé.
4. Décodage (via genome_decoder) : lecture du fichier .dna et restitution des données originales

Chaque étape du pipeline est configurable, permettant d'ajuster le comportement du compressuer selon la nature des données ou les besoins en performance.

3.3 Interfaces et intégration

L'outil dispose d'une interface en ligne de commande (CLI) légère et intuitive, adaptée aux environnements de production ou de test. Le code source est structuré selon les bonnes pratiques du développement open source :

- séparation des modules
- présence de test unitaire
- documentation intégrée (docstrings mutlilingues FR/EN)

3.4 Jeux de données utilisés

Pour les tests et benchmarks initiaux, plusieurs types de jeux de données ont été employés :

- Journaux applicatifs (logs de serveurs, traces systèmes, logs réseau)
- Textes bruts (fichiers de documentation, scripts)
- Données semi-structurees (JON, CSV, format de traces techniques)

Cette diversité permet d'évaluer la robustesse et la généricité du système sur des cas d'usage concrets.

4 Performance et Benchmarks

L'évaluation des performances de GENOME_COMPRESSOR repose sur des tests empirique visant à mesurer son taux de compression, sa vitesse de traitement, ainsi que sa résilience face à différents types de données. Les comparaisons sont effectuées avec des outils de compression généralistes largement utilisés, afin de situer le système dans un contexte applicatif réel.

4.1 Méthodologie d'évaluation

Chaque test suit une procédure standardisée :

1. Donnée d'entrée selectionnée (log applicatif, texte brute, etc.)
2. Compression via GENOME_COMPRESSOR
3. Compression via des outis classiques (gzip, bzip2, zstd)

4. Mesure de la taille en sortie, du temps de traitement, et vérification de la fidélité après décompressions

Les test ont été réalisés sur un poste de travail standard (processuer core i5 5e génération , 8 Go de RAM) sous l'environnement Linux

4.2 Résultats comapratifs

Les résultats obtenus montrent que GENOME_COMPRESSOR offre des taux de compression compétitifs sur certains types de données semi-structurées et textuelles, particulièrement sur les journaux applicatifs contenant des motifs répétitifs.

Type de donnée	**Taille initiale**	**gzip**	**bzip2**	**GENOME_COMPRESSOR**
Logs Apache (2 Mo)	2,00 Mo	0,55 Mo	0,45 Mo	0,42 Mo
JSON structuré (1 Mo)	1,00 Mo	0,35 Mo	0,30 Mo	0,28 Mo
Script Python (100 ko)	0,10 Mo	0,04 Mo	0,03 Mo	0,03 Mo

Table 1 : Résultats comparatifs de compression

Ces résultats sont encourageants, bien qu'ils dépendent fortement de la nature des données et de leur régularité structurelle. Sur des fichier binaires ou tres bruités, les performances peuvent être moindres, le compresseur ayant été initalement optimisé pour des données à motifs.

4.3 Limite observées

À ce stade, plusieurs limites sont identifiées :

- Le temps de traitement est plus élevé que les compresseurs classique pour de très gros volumes (>100 Mo), en raison du surcoût algorithmique lié à l'encodage bio-inspiré
- Le système n'est pas encore multithreadé, ce qui limite son usage en contexte haute performance
- L'algorithme de mutation necessite une calibration plus fine pour certaine jeux de données bruitées.

Ces limites font l'objet d'un suivi actif dans la feuille de route du projet, avec des optimisations en cours de développement

5 Sécurité et extensions prévues

L'approche actuelle le GENOME_COMPRESSOR repose avant tout sur la compression bio-inspirée. Toutefois, la sécurité des données constitue une axe statégique de développement pour les futures versions du projet. Des fonctionnalités complémentaires sont à l'étude afin d'étendre les capacités du système à des contextes d'usage nécessitant une meilleure confidentialité, une traçabilité renforcée ou une intégrité accrue des données.

5.1 Perspectives de renforcement de la sécurité

L'architecture modulaire de `GENOME_COMPRESSOR` ouvre la voie à diverses pistes d'enrichisssement fonctionnel, notamment en matière de confidentialité et d'intégrité des données. Des travaux exploratoires sont en cours afin d'étudier l'intégration de mécanismes inspirés de la cryptographie légère, dans un cadre cohérent avec la bio-inspiration du système.

5.2 Adaptabilité à des environnements à haut niveau d'exigence

À plus long terme, le système pourrait être adapté à des contextes d'usage présentant des contraintes accrues, qu'il s'agisse de conformité, de traçqbilité ou d'automatisation. Ces perspectives soulèvent des questions de recherche touchant à l'architecture, à la robustesse et à l'intégration dans des environnements complexes, qui feront l'objet d'études futures.

5.3 Statut actuel et travaux en cours

La version actuelle de `GENOME_COMPRESSOR` se concentre sur les mécanismes de compression bio-inspirés, disponibles sous licence libre et accessibles à la communauté scientifique et technique, Elle constitue une base expérimentale sur laquelle différentes axes de recherches peuvent être explorés, notamment en matière de modularité, de sécurité et d'adaptabilité.

Des expérimentations sont en cours afin d'évaluer l'intégration de fonctionnalités supplémentaires, dans le cadre d'une démarche ouverte, itérative et collaborative.

5.4 Statut actuel et perspectives

La version actuelle de `GENOME_COMPRESSOR`, publiée sous licence libre, implémente un ensemble de modules bio-inspirés pour la compression de données. Elle constitue

une première étape expérimentale, conçus pour être facilement extensible et testable dans différents contextes.

Les travaux en cours portent sur l'évaluation de nouvelles fonctionnalités, en particulier dans les domaines de la securité, de l'optimisation et de l'intégration dans des chaînes de traitement automatisées. Le projet reste ouvert à la collaboration et à l'exploration de cas d'usage variés. presonnalisation et l'hébergement sécurisé

6 Positionnement Open Science

Le projet GENOME_COMPRESSOR s'inscrit pleinement dans les principes de la science ouverte, en assurant la transparence, l'accessibilité et la réutilisabilité des resultats produits. Ce positionnement se manifeste à plusieurs niveaux, tant sur le plan technique que sur le plans éthique

6.1 Code source libre et accessible

L'intégralité du code source est publiquement disponible sous licence MIT. Cette démarche vise à encourager la transparence, la reproductibilité scientifique et la collaboration ouverte. Le dépôt GitHub associé permet un accès direct à l'implementation, ainsi qu'à la documentation technique du projet.

Ce choix de licence reflète une volonté claire de contribution à la communauté scientifique et open source.

6.2 Documentation ouverte et multilingue

Le projet est accompagné d'une documentation complète, conçue pour faciliter son adoption par un large public :

- Manuel d'utilisation en ligne (FR/EN)
- Commentaires de code normalisés (PEP 257)
- Tutoriels pour développeurs chercheurs et ingénieurs

L'effort de multilinguisme permet de lever les barrières linguistiques, notament dans la communautés francophones, nord-africaines ou sud-américaines.

6.3 Réutilisabilité scientifique

Les composants du projet sont conçus pour être modulaires, interopérables et adaptables à divers cas d'usage. Cette modularité facilite :

- la reproduction des expérience décrite dans l'article

- l'intégration dans d'autre outils de recherche ou projets open source
- la réutilisation dans des contexte éducatif ou industriels

De plus, les jeux de données utilisés pour les tests (logs applicatifs, textes généraux) seront rendus disponibles sous des licences compatibles, dans le respect des règles éthiques et de confidentialité.

7 Conclusion et perspectives

Le projet `GENOME_COMPRESSOR` propose une approche originale de la compression de données, s'inspirant des mécanisme biologiques d'encodage de l'ADN. Conçu autour d'une architecture modulaire, il vise à répondre aux enjeux contemporains liés à la gestion de volumes croissants de données, tout en conservant une forte dimension expérimentale et réutilisable.

La version open source actuellement disponible intègre les principaux modules du pipeline : détection de motifs, encodage génétique, traitement des mutations, compression et décodage réversible. Elle permet déjà de traiter divers types de données textuelles avec des résultats prometteurs en termes de taux de compression et de robustesse face à la variabilité des sources.

Des travaux complémentaires sont en cours afin d'explorer l'intégration de mécanismes de sécurité, l'optimisation des performances, ainsi que l'adaptabilité du système à différents contextes applicatifs, qu'ils soient scientifiques ou opérationnels. Ces axes restent ouverts à la collaboration et à l'expérimentation.

La communauté est encouragée à participer, via des contributions au code, des retours d'expérience, ou des partenariats de recherche. Le projet s'inscrit dans une démarche ouverte et écolutive, en lien avec les dynamiques actuelles de science reproductible et d'innovation bio-inspirée.

Références

1. Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. IEEE Transactions on Information Theory, 23(3), 337-343
2. Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems. Science, 266(5187), 1021–1024.
3. Cantu-Paz, E. (2000). Efficient and Accurate Parallel Genetic Algorithms. Springer.
4. Clelland, C. T., Risca, V., & Bancroft, C. (1999). Hiding messages in DNA microdots. Nature, 399(6736), 533–534.

5. Rivest, R. L., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), 120–126.

6. Li, M., & Vitányi, P. (2008). An Introduction to Kolmogorov Complexity and Its Applications. Springer.

7. Rizzo, L., & Vicisano, L. (1997). RLC : A reliable multicast protocol for wireless environments. In IEEE WCNC.

8. Manel, T., et al. (2023). Compression de logs sécurisée par méthodes bio-inspirées : étude comparative. Conférence AFIA (Association Français pour l'intelligence Artificielle) , France.

9. OpenSSL Project. (2022). OpenSSL : Cryptography and SSL/TLS Toolkit. Consulté sur : https ://www.openssl.org

10. GENOME_COMPRESSOR - Projet open source. (2025). Code source disponible sur : https ://github.com/Skillborn/GENOME_COMPRESSOR