The Journal of Open Source Software

# DNA Data Storage: Error Correction with Dynamic Mapping Rule

**Jessica Schindler[1*], Omar El Menuawy[1*], Tamara Hadzic[1], Lena Wiese[1¶], and Babak Saremi[1]**

**1** Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), Hannover, Germany ¶ Corresponding author * These authors contributed equally.

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:**

**Submitted:** 26 February 2024
**Published:** unpublished



## Summary

The amount of generated data has multiplied rapidly in recent years. Conventional storage methods will eventually reach their limits and may no longer be sufficient to store the amount of data. Therefore, a new long-term storage method that can primarily store archival data is needed in order to be able to store the increasing amount of data in the coming years. DNA offers the advantages of long-term storage, high storage density and low energetic cost, causing the scientific community to attempt to make DNA usable as a storage method. This program contains the coding and error correction procedure of our self-developed "dynamic mapping rule" DMR method for storing data in DNA. The error correction of the method is based on the comparison of the faulty sequence with the mapping scheme and a level-based correction, which specifically addresses the known locations where the mapping does not match. By combining this with the Reed Solomon (RS) code, the valid segments of the DMR correction can be confirmed. Simulations have shown that the correction with the DMR scheme works better than the correction with RS alone.

## Statement of need

It is estimated that the amount of data will be up to 180 zettabytes in 2025, an increase of up to 23% from 2020 to 2025 (IDC, 2021). Based on these forecasts, a new storage option is being sought, whereby the storage of data in DNA is being considered. DNA offers the advantages of extremely high information density and longevity, enabling space-saving and resource-conserving long-term storage. To store data in DNA, it must first be translated into DNA using a transcription scheme so that the encoded sequence can then be synthesized and sequenced. During synthesis and sequencing, errors can occur, therefore they have to be corrected before decoding to ensure that an error-free output file can be restored.

The popular error-correcting coding system is Reed Solomon (RS). It is used in a wide range of applications in digital communications and data storage and find use in mass storage devices (hard disk drives, DVD, barcode tags), wireless and mobile communications units, satellite links, digital TV, digital video broadcasting (DVB), and modem technologies like xDSL (Shrivastava & Singh, 2013). To increase the efficiency of the correction in data storage applications, the RS code is used in combination with the own correction method. For example, Erlich and Zielinski (Erlich & Zielinski, 2017), Press et. al (Press et al., 2020) and Blawat et. al (Blawat et al., 2016) use RS within their developed correction methods.

For efficient decoding of data, it is necessary to determine an optimal correction method. This requires different methods to be developed, tested and compared with each other. For this reason, we developed the "dynamic mapping rule" (DMR) method, which provides coding

and decoding including error correction of data. The method is based on a mapping table, which is used for the translation and also gives the dynamic mapping rule its name. With the help of this mapping table, a self-correction of the DNA sequence is achieved. For encoding, the input sequence was divided into segments, translated and reassembled using defined DNA sequences known as spacers. The spacers are placed between the individual segments so that the segments can be separated and corrected during decoding using the spacer sequences. The selection of the segment size is variable, but it is kept small so that the correction with DMR works better and the runtime is reduced. In addition, the coding was combined with RS so that the correction with the DMR scheme can be validated with the help of RS and thus leads to a better correction than the stand-alone use of DMR or RS. Because RS can only correct a certain number of errors, which depends on the encoded error correction symbols, it is not able to correct errors if the maximum error limit is exceeded. However, the use of DMR enables the correction of error overflows and ensures that the number of errors remains within the correction limits of RS, thus restoring RS functionality. Furthermore, the use of the DMR scheme offers the possibility to localize incorrect areas of the DNA by comparison with the mapping possibilities and thus carry out targeted corrective procedures at the affected sites. This should make it possible to correct substitution as well as insertion and deletion errors.

The correction in the DMR scheme is executed with different designed levels, which proceed one after the other and try out different methods for the correction. In the first level, for example, the incorrect sections are determined and an attempt is made to replace incorrect bases at precisely these points and thus repair the errors. To do this, the mapping table is consulted and all possible combinations for the previous and next 2-mers are searched for. Of these identified 2-mers, only those that show at least 1 matching base to the faulty 2-mere are tested. The possible sequences are checked using the RS code to correct them. This step determines whether the possibilities found are correct or not. The subsequent levels are designed in such a way that more and more possibilities are found if no correction has been made in the previous level. For the correction, the levels must be adapted and expanded accordingly so that an optimal correction can be achieved. Figure 1 shows the schematic illustration of the described DMR correction.

The current code was developed internally as part of the BIOSYNTH project and is now being published as open software. The current designed levels are only suitable for the correction of substitution errors, but it is possible to extend them for the correction of insertion and deletion errors. The developed DMR method was compared with the RS method by encoding the small Fraunhofer logo with the segmented bit array encoding method using different amounts of error-correcting symbols (example in Figure 1). Substitution errors were then added to the encoded DNA and decoded. A total of 20 trials for DMR and 50 trials for RS were performed per setting. Figure 2 shows the results of this comparison and shows that the correction with the DMR scheme produces higher edit distances than the correction with RS, which proves the more effective correction of substitution errors.

The Journal of Open Source Software

## Figures

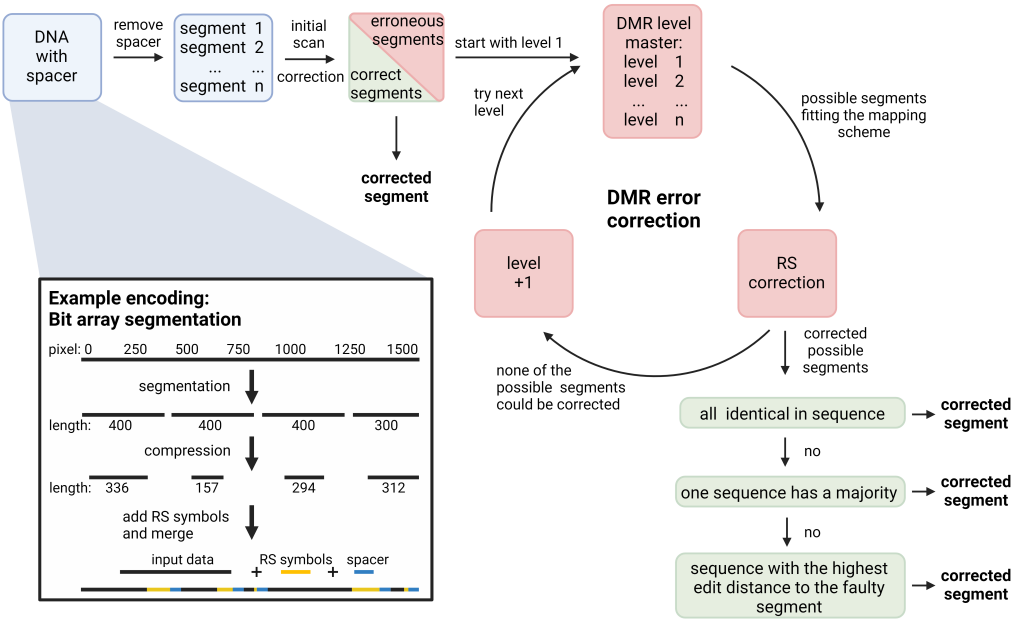


**Figure 1:** Schematic illustration of the dmr error correction. Firstly, the spacers are removed from the DNA sequence and the segments are initially scanned for correct and incorrect sequences. The incorrect segments then enter the DMR correction process. The levels that are passed through attempt to create segments that match the mapping scheme, these segments are then verified with RS. If the check with RS does not work, an attempt is made to correct the incorrect segment with the next level. If the check works, all corrected possible segments are analysed. If they are all the same or one segment has the majority, it is accepted as corrected. If none of the segments stand out by majority, the segment with the greatest similarity to the incorrect segment is assumed to be corrected. Created with BioRender.com

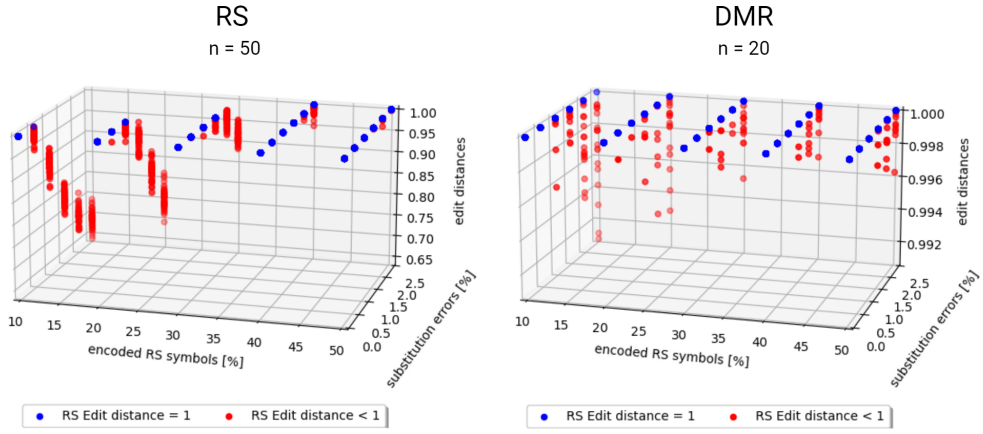


**Figure 2:** Comparison of DMR and RS correction with different settings and error rates of the bit array segmentation coding method. The simulation of the DMR correction is shown on the left and was performed with 20 trials. The simulation with RS is shown on the right and was performed with 50 trials. In both diagrams, fully corrected images with an edit distance of 1 are shown in blue and erroneous images with a smaller edit distance are shown in red. The coding was carried out with the help of a bit array segmentation. Here, the image is first read in and then the pixels are divided into segments. The individual segments are then compressed and translated.

## Acknowledgements

This work was funded by the Fraunhofer PREPARE Programm and developed within the BIOSYNTH project under the project number 40-03168-2420 as part of the data coding research.

## Author Contributions

J.S., B.S. and L.W. wrote the paper. J.S. created the figures, helped with the creation of the main functions, enhanced the error correction, designed and performed the simulation. O.M. has written the base code of the DMR mapping and error correction. T.H. helped with the creation of the main functions and wrote the code for the edit distance calculation and error simulation. B.S., L.W. and O.M. conceptualized the DMR mapping. All authors reviewed the manuscript.

## References

Blawat, M., Gaedke, K., Hütter, I., Chen, X.-M., Turczyk, B., Inverso, S., Pruitt, B. W., & Church, G. M. (2016). Forward error correction for DNA data storage. *Procedia Computer Science*, *80*, 1011–1022. https://doi.org/10.1016/j.procs.2016.05.398

Erlich, Y., & Zielinski, D. (2017). DNA fountain enables a robust and efficient storage architecture. *Science (New York, N.Y.)*, *355*(6328), 950–954. https://doi.org/10.1126/science.aaj2038

IDC. (2021). *Worldwide global DataSphere forecast, 2021–2025: The world keeps creating more data — now, what do we do with it all?* (IDC, Ed.; IDC Doc #US46410421). https://www.marketresearch.com/IDC-v2477/Worldwide-Global-DataSphere-Forecast-Keeps-14315439/

Press, W. H., Hawkins, J. A., Jones, S. K., Schaub, J. M., & Finkelstein, I. J. (2020). HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(31), 18489–18496. https://doi.org/10.1073/pnas.2004821117

Shrivastava, P., & Singh, U. P. (2013). Error detection and correction using reed solomon codes. *International Journal of Advanced Research in Computer Science and Software Engineering*, *Volume 3*. https://api.semanticscholar.org/CorpusID:18920204