A New Data Modeling Approach for Alignment-free Biological Applications

Diogo Munaro Vieira[1], Elvismary Molina de Armas[1],
Maria L. G. Jaramillo [1], Marcos Catanho[2], Antonio B. Miranda[2],
Edward Hermann Haeusler[1], Sérgio Lifschitz[1]

[1]Informatics Dept., Pontifical Catholic Univ. of Rio de Janeiro (PUC-Rio), Brazil

[2]Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, Brazil

{dvieira, earmas, mjaramillo, hermann, sergio}@inf.puc-rio.br,

{mcatanho, antonio.miranda}@ioc.fiocruz.br

Abstract. *Finding homologous proteins and grouping them are tasks of utmost importance in biology, which currently rely on tools based on information from these proteins' DNA or amino acid sequences. These tasks require identifying evolutionary patterns that are challenging to obtain automatically using traditional methods. This work proposes a data modeling approach to leverage evolutionary patterns in homology searching, ranking, and clustering tasks through an alignment-free process using image similarity algorithms. This strategy is valuable even for distant homologs and contributes to data privacy and security.*

Resumo. *Encontrar proteínas homólogas e agrupá-las são tarefas de extrema importância para a biologia, que atualmente conta com ferramentas baseadas em informações do DNA ou das sequências de aminoácidos dessas proteínas. Essas tarefas exigem a identificação de padrões evolutivos que são difíceis de obter automaticamente usando métodos tradicionais. Este trabalho propõe uma abordagem de modelagem de dados para alavancar padrões evolutivos em tarefas de busca, classificação e agrupamento de homólogos por meio de um processo alignment-free usando algoritmos de similaridade de imagem. Essa estratégia é valiosa mesmo para homólogos distantes e contribui para a privacidade e segurança dos dados.*

1. Introduction

Computers have become an indispensable aspect of modern society, profoundly impacting every facet of our lives. Their remarkable capacity to process and store immense volumes of information is a key attribute that renders computers highly valuable. This ability to handle vast datasets has fostered the development of a diverse array of applications and technologies. However, for computers to effectively process data, it is crucial to organize and format it in a manner that the machines can comprehend. This necessitates understanding the nature of the data and establishing connections, achieved through the use of conceptual and logical models. Once these initial steps are accomplished, the data can be translated into a practical representation through a physical model, facilitating its utilization in real-world scenarios.

Bioinformatics involves the processing and analysis of vast amounts of data from various sources, including DNA and protein sequences, gene expression profiles, and metabolic pathways. The most commonly used data format in bioinformatics for homology study is text representation through FASTA files [Mills 2014]. It provides a simple and ready-to-use way to store sequence information and annotations of DNA, RNA, or protein molecules, which can be visualized in any text editor application and is easy to process for most programs. Relational information can also be stored in similar text files representing for instance nucleotide or amino acid sequence alignments (ClustalW, MEGA, MSF, etc), phylogenetic relationships (NEXUS, NEWICK, PHILIP, etc) [Leonard et al. 2006], or 3D molecular structure (PDB) [Bernstein et al. 1977].

The dot plot is a valuable tool for simultaneously assessing sequence and structural similarities between homologous sequences. It involves comparing a sequence to itself (self-comparing) or to another sequence (inter-comparing) generating 2D graphs with matches between sequences [Gibbs and Mcintyre 1970]. On the other hand, dot plots provide visual representations of evolutive processes (substitutions, insertions, and deletions) and internal structural information (direct and inverted repeats) that may not be evident in the data representations discussed earlier. However, the current approach for extracting and utilizing this information primarily relies on manual observation. Another aspect of the dot plot that has not been thoroughly explored is its potential for data masking. Data masking techniques have been extensively studied in the field of preserving the privacy of biological data, particularly in medical contexts, and they provide a means to ensure data security without compromising access to information [Siddartha and Ravikumar 2019, Siddartha and Ravikumar 2020].

There are quite conventional data modeling approaches for DNA [Lifschitz et al. 2022, Bilotta et al. 2019], but we encounter more difficulty when dealing with representations of evolutive processes. An interesting data representation approach has considered sequences as images through the *Chaos Game Representation* (CGR), where the DNA sequence is translated into an image using a mathematical function that places each nucleotide at a specific location in the image based on its position in the DNA. CGR and its compact version based on letter frequencies (FCGR) have various applications in biology [Löchel and Heider 2021]. Despite being able to represent many interesting characteristics of sequences, CGR still have issues detecting evolutive processes [Kania and Sarapata 2021], and it does not allow for gains in explainability and security.

In this work, we will focus on molecular homology, particularly orthologous proteins, which descending from a common evolutionary ancestor, tend to perform the same function in different species [Fitch 1970]. This work proposes a novel physical model inspired by dot plots for the data modeling approach to leverage evolutionary patterns in homology searching, ranking, and clustering tasks through an alignment-free process using image similarity algorithms. This strategy is valuable even for distant homologs and contributes to data privacy and security integrating with bioinformatics workflows.

2. Methods and materials

The dataset, building, and validation process for this work are presented in this section. Validation measures to assess the performance are described for homology clustering and searching too.

2.1. Dataset

The dataset comprises DNA sequences of protein-coding genes of Globin family, obtained from Ensembl [1], including five distinct groups of homologous proteins (Hemoglobin$_\beta$, Myoglobin, Neuroglobin, Cytoglobin, and Androglobin), with 15 species (Chlorocebus Sabaeus, Otolemur Garnettii, Nomascus Leucogenys, Saimiri Boliviensis, Prolemur Simus, Aotus Nancymaae, Gorilla Gorilla, Pan Troglodytes, Rhinopithecus Roxellana, Rhinopithecus Bieti, Pan Paniscus, Mandrillus Leucophaeus, Carlito Syrichta, Cebus Capucinus, and Pongo Abelii) represented in each group, yielding a total of 75 DNA sequences for analysis). There is a script to extract these five datasets from Ensembl and it's available on project GitHub [2] with name *get_globins.sh*.

Analyzing these homologous datasets, Hemoglobin$_\beta$ and Myoglobin exhibit relatively uniform sequence lengths of 441 and 465 nucleotides respectively, and no standard deviation (σ), while the variation in sequence length gradually increases for Neuroglobin (451.6: $\sigma = 11.43$), Cytoglobin (596.2: $\sigma = 66.25$), and Androglobin (4726.4: $\sigma = 694.56$), successively.

In order to evaluate the effectiveness of the new data model in capturing evolutionary patterns, a combination of synthetic and real data was utilized. Synthetic sequences were generated using the *INDELible* tool [Fletcher and Yang 2009], following the configuration file *indelible.conf* available on the corresponding GitHub repository. A total of 40 random sequences, each consisting of 3000 nucleotides, were created. These sequences were organized into 10 separate lineages, with four sequences within each lineage.

2.2. Image Comparison

The image comparison techniques used in this study rely on methods based on the Human Visual System (HVS) to evaluate whether images are similar using properties that can be perceived by human vision as if we are manually analyzing a dot plot. These methods were employed to identify similarities between generated images and search for homologs using just the images. Blastn (basic local alignment search tool version for nucleotides) [McGinnis and Madden 2004] was used as a baseline to compare the results. A BLAST search provides researchers with the ability to compare a query protein or nucleotide sequence to a database of sequences. It allows for the identification of database sequences that exhibit significant similarity to the query sequence.

Various image similarity algorithms that use HVS were used in this study, ranging from simpler to more complex methods: Universal Quality Index (UQI)[Wang and Bovik 2002], Structural Similarity Index Measure (SSIM)[Wang et al. 2004, Bakurov et al. 2022], and MultiScale Structural Similarity Index Measure (MS-SSIM) [Wang et al. 2003]. SSIM and MS-SSIM were applied with their default settings, without any hyperparameter optimization and UQI adjusted the window to 11 to be comparable with SSIM and MS-SSIM. In the implementation of the similarity matrix using the image comparison algorithms, when two images of different sizes were compared, the smaller one was resized to the same size as the larger image and then compared.

[1] `https://www.ensembl.org/index.html`

[2] `https://github.com/BioBD/DNA2D`

2.3. Proposed Data Model

The proposed new data model for representing DNA is inspired by the 2D dot plots. It represents the internal structures of DNA, allowing for the use of methods to identify patterns of evolutionary characteristics in the images. A common way to represent images is through the RGB (Red, Green, and Blue) channels standard [Plataniotis and Venetsanopoulos 2000]. Each channel represents one entity from our model: direct repeats on the red channel, inverted repeats on the green, and reverse repeats on the blue channel.

Our methodology involved the creation of self-comparison matrices for DNA sequences, where each cell in the matrix was assigned a number to represent the correspondence between identical nucleotides, similar to a regular self-comparison matrix. To enhance the visibility of sections with more matches, we generated images with pixel values ranging from 0 to N nucleotides, where N represents the maximum number of nucleotides followed by a match. Initially, the values in the matrix were set to 1 for matches and 0 for non-matches. The algorithm then calculated the number of consecutive matches in each sequence and assigned the maximum number of matches in a row, denoted as N, as the pixel value in the matrix M. This process is illustrated in Figure 1.

$$M^1 \rightarrow \left|\begin{array}{ccccc} & \mathbf{G} & \mathbf{T} & \mathbf{T} & \mathbf{A} \\ \mathbf{A} & 0 & 0 & 0 & 1 \\ \mathbf{G} & 1 & 0 & 0 & 0 \\ \mathbf{T} & 0 & 1 & 1 & 0 \\ \mathbf{T} & 0 & 1 & 1 & 0 \end{array}\right| \qquad M^N \rightarrow \left|\begin{array}{ccccc} & \mathbf{G} & \mathbf{T} & \mathbf{T} & \mathbf{A} \\ \mathbf{A} & 0 & 0 & 0 & 1 \\ \mathbf{G} & 3 & 0 & 0 & 0 \\ \mathbf{T} & 0 & 3 & 1 & 0 \\ \mathbf{T} & 0 & 1 & 3 & 0 \end{array}\right|$$

Figure 1. Demonstration of matrix filling. First, add the number 1 in all correspondences and then change it to the maximum in each one in a row.

Then these matrix numbers were normalized between 0 and 255, expanding the strength of that color represented according to the greater correspondence of the compared nucleotide windows and keeping the images with the same comparison standard. A pseudo-code with complexity $O(size_1 \times size_2)$ representing how each channel was populated is represented in Algorithm 1, where $size_1$ and $size_2$ are the sizes of the compared sequences.

To store these matrices, we utilized the RGB channels of an image. The comparisons between sequences were made in the following manner: the sequence with itself was stored in the R channel, the sequence with its reverse complement in the G channel, and the sequence with the same inverted in the B channel. This methodology allowed us to incorporate three layers of genetic information into a single image, as depicted in Figure 2 revealing evolutionary processes or patterns. The images were saved in PNG format, which is a lossless compression and keeps the file size small.

2.4. Sequence Searching and Clustering

To establish a comparative benchmark for homologous sequence searching, were employed the Blastn algorithm on a local database that exclusively contained sequences from all datasets described in subsection 2.1. To ensure a fair comparison, we used a $word\_size$ of 11 in the Blastn algorithm, allowing it to search for nucleotide repetitions

Algorithm 1 Pseudo-code for image channel creation by comparing two DNA sequences.

```
function MakeChannel($seq_1, seq_2, pixel_{max} = 255$)
    $size_1 \leftarrow \textsc{Size}(seq_1)$
    $size_2 \leftarrow \textsc{Size}(seq_2)$
    $channel \leftarrow \begin{bmatrix} 0 & 0 & \dots \\ \vdots & \ddots & \\ 0 & & 0 \end{bmatrix}_{(size_1 \times size_2)}$
    for $s_1 = 0, s_1{+}{+}, s_1 \leq size_1$ do
        for $s_2 = 0, s_2{+}{+}, s_2 \leq size_2$ do
            if $seq_1[s_1] = seq_2[s_2]$ then
                $channel[s_1, s_2] = 1$                          ▷ Fill with 1 if match
            end if
        end for
    end for
    $seq_{max} = \textsc{FillDiag}(channel)$           ▷ Fill diagonal with sequential match
    return $channel \times pixel_{max} \div seq_{max}$         ▷ Return normalized channel
end function
```

of 11 nucleotides, similar to the $filter\_size$ parameter used in HVS algorithms, as described in subsection 2.2. For the Blastn search, more parameters were set, including $evalue = 10000$ not filter out results encompassing sequences displaying low similarity scores between them, and $max\_hsps = 1$ to focus only on the best hit. The Blastn Bit score results were obtained for each sequence and ranked based on the match of the first K sequences from the same dataset returned. To evaluate and validate the performance of the information retrieval algorithms in this way, we used the Mean Average Precision (MAP) metric [Alhijawi et al. 2023].

The proposed approach differs from the control in the search phase by transforming the sequences into images before comparing them. These images are then processed by the HVS algorithms to compute a similarity score between each pair. The similarity scores were ranked from most to least similar, and the MAP is calculated from the top K items to validate that only sequences from the same dataset were retrieved.

For clustering, the control was performed using Clustal Omega with default parameters, and a similarity matrix was computed using the image comparison algorithms. Each score in the matrix was inverted $(1 - score)$ to convert it into a dissimilarity matrix for compatibility with the UPGMA algorithm, which is an agglomerative hierarchical clustering method used to build phylogenetic dendrograms [Huelsenbeck 1995]. The Robinson-Foulds (RF) [Robinson and Foulds 1981] distance metric, normalized between 0 and 1, was applied to compare the dendrograms built by the new algorithm against the control. A lower RF metric indicates better results, which means that the obtained dendrogram is similar (or equal) to control.

The experiments and corresponding source code are publicly available on GitHub, ensuring transparency and reproducibility. Additionally, all the data used in the study was versioned and made accessible through Google Drive as part of the project. The complete

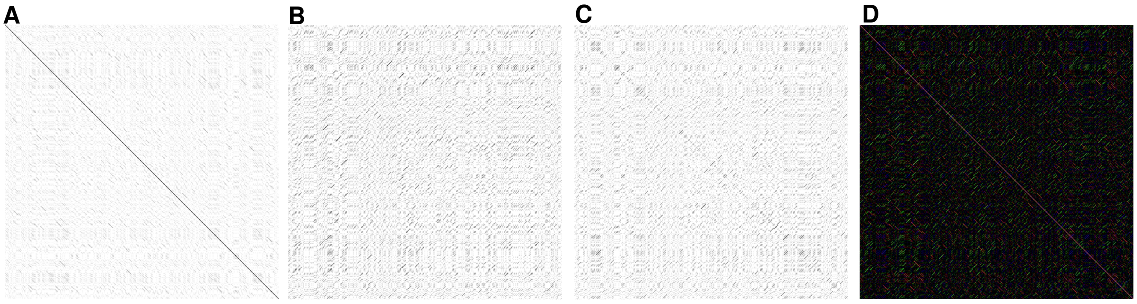

Figure 2. Dot plot of self-comparison matrices between Human Hemoglobin Beta chains made by the proposed methodology. Grayscale values have been reversed for easier viewing. A) only R channel with grayscale sequential auto-comparison; B) G channel only with grayscale reverse complementary; C) B channel only with grayscale inverse; D) mixing between RGB channels forming an image.

Dataset	Identity	# Gaps	Median	Mean	Std	Min	Max
Hemoglobin$_\beta$	0.75	0	0	0	0	0	0
Myoglobin	0.76	0	0	0	0	0	0
Neuroglobin	0.61	2	0	4.4	11.43	0	39
Cytoglobin	0.40	15	66	87.8	66.25	6	306
Androglobin	0.61	15	125	327.6	694.56	50	2906
Indelible	0	40	3302	3302	0	3302	3302

Table 1. Identity, number of gapped sequences and amount of gaps from samples after global alignment with Clustal Omega.

experimental workflow is illustrated in Figure 3.

3. Results

All the results achieved in this study are reproducible and were obtained using Python 3.7+ code with Jupyter Notebooks. By employing these technologies, we conducted homology searching and clustering experiments on all datasets described in subsection 2.1.

3.1. Dataset Insights

In the first experiment, Clustal Omega [Sievers and Higgins 2018] was employed to perform sequence alignment on each dataset in order to extract insights. The obtained data and corresponding statistics can be found in Table 1.

The analysis of these statistics reveals minimal (0.1%) differences in identity between the Myoglobin and Hemoglobin$_\beta$ datasets, indicating the highest similarity datasets with more than 0.7 of identity and without any gaps observed in the multiple sequence alignment (MSA). Despite having a greater variation in sequence length ($\sigma = 694.56$) and a high number of gaps in the MSA (dataset mean of 327.6 gaps), Androglobin exhibits the MSA identity equal to that of Neuroglobin. However, due to its shorter sequence length, Neuroglobin has almost no gaps in the MSA (dataset mean of 4.4 gaps). On the other hand, while Cytoglobin shows a low MSA identity (0.4) with its counterparts, it has fewer gaps (mean only 87.8) compared to Androglobin.

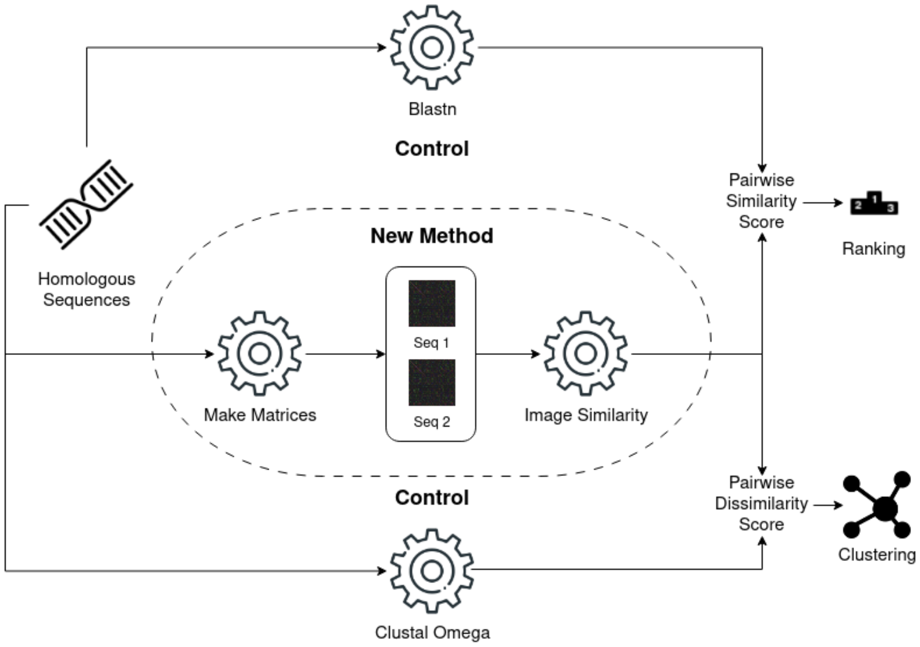

Figure 3. Homology Searching and Clustering experiment workflow. Differences between control (up) and new method (down) from Ranking and control (down) and new method (up) for Clustering.

The Hemoglobin$_\beta$ and Myoglobin dataset comprises two identical sequences across different species. The Neuroglobin set has two pairs of identical sequences, while no identical sequences were found in the Cytoglobin and Androglobin datasets. Synthetic *INDELible* sequences have a fixed number of nucleotides and do not share any MSA identity because of their high dissimilarity. The number of gaps was static in all sequences during the alignment, providing a more stable and diverse dataset. Another notable difference between Globins and synthetic sequences is the Longest Common Subsequence (LCS), a high marker of heredity between DNA sequences [Alsmadi and Nuser 2012, Namiki et al. 2012].

The LCS was obtained by iteratively comparing pairs of sequences within a dataset to determine the longest shared string between them. Subsequently, the median LCS metric was calculated for each dataset by considering all LCS values of that dataset sequences pairwise. It is worth noting that the synthetic sequences exhibited an LCS of only 11 nucleotides, leading to limited evolutionary detection when compared to the LCS values of Globins: 88 for Hemoglobin$_\beta$; 65 for Myoglobin; 77 for Neuroglobin; 113 for Cytoglobin; and 130 for Androglobin.

3.2. Homology Searching

To ensure a fair evaluation of search ranking performance across datasets of varying sizes and properties, we used the control as the main reference. To measure the accuracy of the retrieved results, we employed the MAP metric, which is widely used for evaluating information retrieval systems. Unlike other ranking metrics MAP is particularly useful when it's important to determine whether K items are within the result, because the metric scores items in a binary way (correct or not) based on the ranking order. Table 2 summarizes the MAP scores obtained for all the algorithms and the control, where higher

Dataset	Blast control	UQI	SSIM	MS-SSIM	*K* Number
Hemoglobin$_\beta$	1	1	1	1	15
Myoglobin	1	1	1	1	15
Neuroglobin	1	0.85	0.69	0.74	15
Cytoglobin	1	0.45	0.47	0.56	15
Androglobin	1	0.90	0.21	0.90	15
Indelible	0.66	0.61	0.63	0.98	40

Table 2. MAP results for each dataset searching for K top most similar sequences compared to Blast control.

scores indicate better performance.

MS-SSIM outperformed all other algorithms, including the control, with a MAP result of 0.98 compared to 0.66 for the synthetic *INDELible* dataset. However, when MS-SSIM was excluded, the algorithms, including the control, showed similar performance on this synthetic dataset. Although the synthetic dataset presented challenges for detecting similarities, as expected, MS-SSIM performed well.

The datasets of Hemoglobin$_\beta$ and Myoglobin, as well as *INDELible*, do not have sequences of different sizes, therefore, they achieved the best MAP values (1.0) in the new methodologies with HVS and with the control, returning all the elements correct in the search results. Our HVS implementation just resize images to compare them and with same size sequences we don't need resize getting better results.

Cytoglobin showed low MAP results (around 0.5 for all HVS algorithms) due to the lower sequence identity and significant differences in sequence lengths. On the other hand, despite the varying sizes of Androglobin sequences, reflected in numerous gaps, the results obtained with MS-SSIM supported the presence of similar subsequences to other homologous sequences. MS-SSIM's ability to compare images at multiple scales allowed it to identify similarities even within subimages. This, combined with high LCS in Androglobin as described in subsection 2.1, led to improved results with MS-SSIM (MAP of 0.9). However, it should be possible to achieve similar outcomes by adjusting the SSIM hyperparameters, as UQI performed equal to MS-SSIM without using multiple scales. It is worth noting that SSIM is a more stable version of UQI, with differences in implementation language, stability constants, and the application of a Gaussian filter in the analyzed window [Wang et al. 2004]. In terms of implementation, UQI was not implemented using TensorFlow code, but the version of *Sewar* Python library was used [3].

3.3. Homology Clustering

The MS-SSIM cluster exhibited the best clustering performance with an RF value of 0.16, outperforming UQI and SSIM clusters, which had RF values of 0.19. The crucial aspect in this clustering task is the distinction between distances among sequences from different datasets. This indicates that, while the MS-SSIM algorithm successfully identified similar sequences, it also effectively discriminated between diverse datasets. On the other hand, UQI demonstrated excellent performance in identifying sequences within the same dataset but was not as effective as SSIM in distinguishing between distinct datasets.

[3] `https://pypi.org/project/sewar/`

The dendrogram generated using MS-SSIM closely resembled the Clustal clustering, with only two Neuroglobin sequences positioned incorrectly, as depicted in Figure 4. However, these sequences stand apart from all other groups, preventing misclassification. Upon closer analysis, we discovered that these two sequences, along with two others, possess different sizes within the dataset, and the resizing approach employed by our methodology did not contribute to the desired outcomes. In the same figure, we can observe that the *INDELible* synthetic sequences exhibit variations in the Clustal control, whereas our method places them in more appropriate groups. All remaining sequences align correctly with the control, and even the synthetic sequences exhibit distinct branches for each synthetic lineage.

4. Discussion

The field of DNA data representation is constantly evolving, and our work introduces a novel 2D data model specifically designed for homology analysis. We have developed an innovative alignment-free approach that yields promising results in homology searching, ranking, and clustering tasks. This novel methodology contributes to data representation and analyses in the field of bioinformatics.

One key advantage of our approach is its independence from an evolutionary model. Instead of relying on predefined models, we focus on comparing the impact of various evolutionary events on groups of homologous sequences. This allows for a comprehensive analysis of the entire sequence, enabling researchers to gain insights that may be missed when examining only selected residues. By taking this holistic approach, our methodology opens up new possibilities for understanding evolutionary patterns at a deeper level.

The Indelible dataset simulates distant homologs and the new data model combined with the comparison methodology presented here is a potential method to detect distant homologs instead of using only the technique with the secondary structure as usually done [Ginalski et al. 2004, Ginalski et al. 2003], but more research is needed on this field to assert that.

In addition to this possible application with distant homologs, the new DNA representation hides nucleotide data, making it unnecessary to store the actual genome data, but simply the representations of evolutionary patterns such as direct and inverse repetitions and substitutions which are crucial for understanding genetic relationships. DNA through representations in the R, G and B channels includes a security layer for data masking on the actual organism data, ensuring the privacy of individuals' genomic information, a critical concern in today's data-driven world.

Along with data representation, the HVS image analysis technique was introduced in bioinformatics as a measure of distance in a search algorithm. By leveraging the visual patterns present in DNA sequences, we can measure distance and similarity using an innovative approach. This method contributes not only to the field of bioinformatics but also to the domain of information retrieval, image analysis and pattern recognition.

We chose HVS algorithms based on the structural information they provide, highlighting the strong interdependencies among pixels, particularly when they are spatially proximate. This interdependence bears a striking resemblance to the interdependence

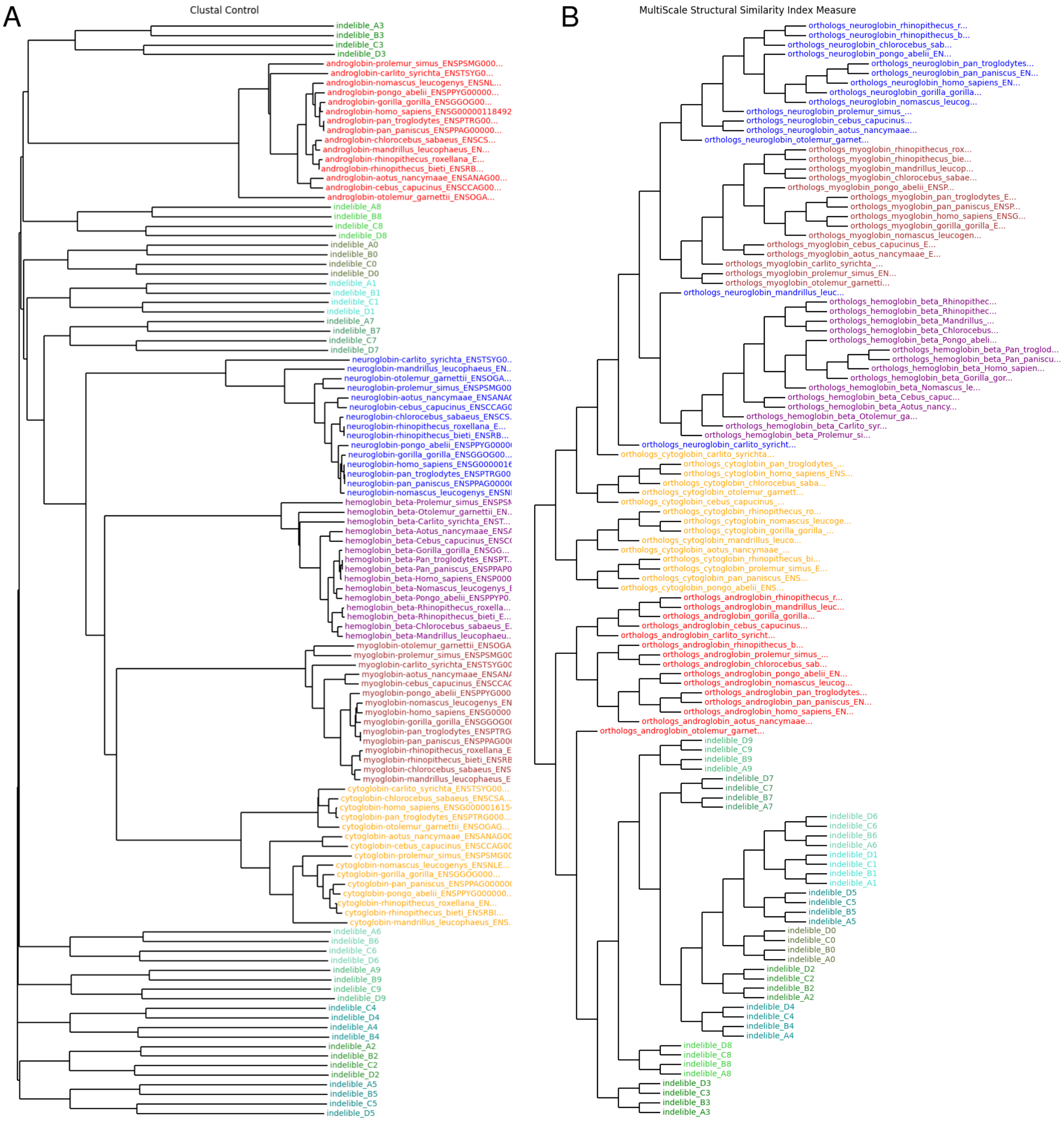
A
Clustal Control
indelible_A3
indelible_B3
indelible_C3
indelible_D3
androglobin-prolemur_simus_ENSPSMG000...
androglobin-carlito_syrichta_ENSTSYG0...
androglobin-nomascus_leucogenys_ENSNL...
androglobin-pongo_abelii_ENSPPYG00000...
androglobin-gorilla_gorilla_ENSGGOG00...
androglobin-homo_sapiens_ENSG00000118492
androglobin-pan_troglodytes_ENSPTRG00...
androglobin-pan_paniscus_ENSPPAG00000...
androglobin-chlorocebus_sabaeus_ENSCS...
androglobin-mandrillus_leucophaeus_EN...
androglobin-rhinopithecus_roxellana_E...
androglobin-rhinopithecus_bieti_ENSRB...
androglobin-aotus_nancymaae_ENSANAG00...
androglobin-cebus_capucinus_ENSCCAG00...
androglobin-otolemur_garnettii_ENSOGA...
indelible_A8
indelible_B8
indelible_C8
indelible_D8
indelible_A0
indelible_B0
indelible_C0
indelible_D0
indelible_A1
indelible_B1
indelible_C1
indelible_D1
indelible_A7
indelible_B7
indelible_C7
indelible_D7
neuroglobin-carlito_syrichta_ENSTSYG0..
neuroglobin-mandrillus_leucophaeus_EN..
neuroglobin-otolemur_garnettii_ENSOGA...
neuroglobin-prolemur_simus_ENSPSMG00
neuroglobin-aotus_nancymaae_ENSANA
neuroglobin-cebus_capucinus_ENSCCAG0
neuroglobin-chlorocebus_sabaeus_ENSCS.
neuroglobin-rhinopithecus_roxellana_E...
neuroglobin-rhinopithecus_bieti_ENSRB...
neuroglobin-pongo_abelii_ENSPPYG0000
neuroglobin-gorilla_gorilla_ENSGGOG00...
neuroglobin-homo_sapiens_ENSG000001
neuroglobin-pan_troglodytes_ENSPTRG00
neuroglobin-pan_paniscus_ENSPPAG0000
neuroglobin-nomascus_leucogenys_ENSN
hemoglobin_beta-Prolemur_simus_ENSPS
hemoglobin_beta-Otolemur_garnettii_EN..
hemoglobin_beta-Carlito_syrichta_ENST...
hemoglobin_beta-Aotus_nancymaae_ENSA
hemoglobin_beta-Cebus_capucinus_ENSC
hemoglobin_beta-Gorilla_gorilla_ENSGG...
hemoglobin_beta-Pan_troglodytes_ENSPT..
hemoglobin_beta-Pan_paniscus_ENSPPAP0
hemoglobin_beta-Homo_sapiens_ENSP000
hemoglobin_beta-Nomascus_leucogenys_
hemoglobin_beta-Pongo_abelii_ENSPPYP0.
hemoglobin_beta-Rhinopithecus_roxella...
hemoglobin_beta-Rhinopithecus_bieti_E...
hemoglobin_beta-Chlorocebus_sabaeus_E.
hemoglobin_beta-Mandrillus_leucophaeu..
myoglobin-otolemur_garnettii_ENSOGA
myoglobin-prolemur_simus_ENSPSMG00
myoglobin-carlito_syrichta_ENSTSYG00
myoglobin-aotus_nancymaae_ENSANA
myoglobin-cebus_capucinus_ENSCCAC
myoglobin-pongo_abelii_ENSPPYG000
myoglobin-nomascus_leucogenys_EN
myoglobin-homo_sapiens_ENSG0000
myoglobin-gorilla_gorilla_ENSGGOG0
myoglobin-pan_troglodytes_ENSPTRG
myoglobin-pan_paniscus_ENSPPAG00
myoglobin-rhinopithecus_roxellana_E
myoglobin-rhinopithecus_bieti_ENSRB
myoglobin-chlorocebus_sabaeus_ENS
myoglobin-mandrillus_leucophaeus_E
cytoglobin-carlito_syrichta_ENSTSYG00...
cytoglobin-chlorocebus_sabaeus_ENSCSA...
cytoglobin-homo_sapiens_ENSG00000161540
cytoglobin-pan_troglodytes_ENSPTRG000...
cytoglobin-otolemur_garnettii_ENSOGAG...
cytoglobin-aotus_nancymaae_ENSANAG00
cytoglobin-cebus_capucinus_ENSCCAG00
cytoglobin-prolemur_simus_ENSPSMG00
cytoglobin-nomascus_leucogenys_ENSNLE...
cytoglobin-gorilla_gorilla_ENSGGOG000...
cytoglobin-pan_paniscus_ENSPPAG00000
cytoglobin-pongo_abelii_ENSPPYG000000..
cytoglobin-rhinopithecus_roxellana_EN...
cytoglobin-rhinopithecus_bieti_ENSRBI...
cytoglobin-mandrillus_leucophaeus_ENS.
indelible_A6
indelible_B6
indelible_C6
indelible_D6
indelible_A9
indelible_B9
indelible_C9
indelible_D9
indelible_C4
indelible_D4
indelible_A4
indelible_B4
indelible_A2
indelible_B2
indelible_C2
indelible_D2
indelible_A5
indelible_B5
indelible_C5
indelible_D5
B
MultiScale Structural Similarity Index Measure
orthologs_neuroglobin_rhinopithecus_r...
orthologs_neuroglobin_rhinopithecus_b...
orthologs_neuroglobin_chlorocebus_sab...
orthologs_neuroglobin_pongo_abelii_EN...
orthologs_neuroglobin_pan_troglodytes...
orthologs_neuroglobin_pan_paniscus_EN...
orthologs_neuroglobin_homo_sapiens_EN...
orthologs_neuroglobin_gorilla_gorilla...
orthologs_neuroglobin_nomascus_leucog...
orthologs_neuroglobin_prolemur_simus_...
orthologs_neuroglobin_cebus_capucinus...
orthologs_neuroglobin_aotus_nancymaae...
orthologs_neuroglobin_otolemur_garnet...
orthologs_myoglobin_rhinopithecus_rox...
orthologs_myoglobin_rhinopithecus_bie...
orthologs_myoglobin_mandrillus_leucop...
orthologs_myoglobin_chlorocebus_sabae...
orthologs_myoglobin_pongo_abelii_ENSP...
orthologs_myoglobin_pan_troglodytes_E...
orthologs_myoglobin_pan_paniscus_ENSP...
orthologs_myoglobin_homo_sapiens_ENSG...
orthologs_myoglobin_gorilla_gorilla_E...
orthologs_myoglobin_nomascus_leucogen...
orthologs_myoglobin_cebus_capucinus_E...
orthologs_myoglobin_aotus_nancymaae_E...
orthologs_myoglobin_carlito_syrichta_...
orthologs_myoglobin_prolemur_simus_EN...
orthologs_myoglobin_otolemur_garnetti...
orthologs_neuroglobin_mandrillus_leuc...
orthologs_hemoglobin_beta_Rhinopithec...
orthologs_hemoglobin_beta_Rhinopithec...
orthologs_hemoglobin_beta_Mandrillus_...
orthologs_hemoglobin_beta_Chlorocebus...
orthologs_hemoglobin_beta_Pongo_abeli...
orthologs_hemoglobin_beta_Pan_troglod...
orthologs_hemoglobin_beta_Pan_paniscu...
orthologs_hemoglobin_beta_Homo_sapien...
orthologs_hemoglobin_beta_Gorilla_gor...
orthologs_hemoglobin_beta_Nomascus_le...
orthologs_hemoglobin_beta_Cebus_capuc...
orthologs_hemoglobin_beta_Aotus_nancy...
orthologs_hemoglobin_beta_Otolemur_ga...
orthologs_hemoglobin_beta_Carlito_syr...
orthologs_hemoglobin_beta_Prolemur_si...
orthologs_neuroglobin_carlito_syricht...
orthologs_cytoglobin_carlito_syrichta...
orthologs_cytoglobin_pan_troglodytes_...
orthologs_cytoglobin_homo_sapiens_ENS...
orthologs_cytoglobin_chlorocebus_saba...
orthologs_cytoglobin_otolemur_garnett...
orthologs_cytoglobin_cebus_capucinus_...
orthologs_cytoglobin_rhinopithecus_ro...
orthologs_cytoglobin_nomascus_leucoge...
orthologs_cytoglobin_gorilla_gorilla_...
orthologs_cytoglobin_mandrillus_leuco...
orthologs_cytoglobin_aotus_nancymaae_...
orthologs_cytoglobin_rhinopithecus_bi...
orthologs_cytoglobin_prolemur_simus_E...
orthologs_cytoglobin_pan_paniscus_ENS...
orthologs_cytoglobin_pongo_abelii_ENS...
orthologs_androglobin_rhinopithecus_r...
orthologs_androglobin_mandrillus_leuc...
orthologs_androglobin_gorilla_gorilla...
orthologs_androglobin_cebus_capucinus...
orthologs_androglobin_carlito_syricht...
orthologs_androglobin_rhinopithecus_b...
orthologs_androglobin_prolemur_simus_...
orthologs_androglobin_chlorocebus_sab...
orthologs_androglobin_pongo_abelii_EN...
orthologs_androglobin_nomascus_leucog...
orthologs_androglobin_pan_troglodytes...
orthologs_androglobin_pan_paniscus_EN...
orthologs_androglobin_homo_sapiens_EN...
orthologs_androglobin_aotus_nancymaae...
orthologs_androglobin_otolemur_garnet...
indelible_D9
indelible_C9
indelible_B9
indelible_A9
indelible_D7
indelible_C7
indelible_B7
indelible_A7
indelible_D6
indelible_C6
indelible_B6
indelible_A6
indelible_D1
indelible_C1
indelible_B1
indelible_A1
indelible_D5
indelible_C5
indelible_B5
indelible_A5
indelible_D0
indelible_C0
indelible_B0
indelible_A0
indelible_D2
indelible_C2
indelible_B2
indelible_A2
indelible_D4
indelible_C4
indelible_B4
indelible_A4
indelible_D8
indelible_C8
indelible_B8
indelible_A8
indelible_D3
indelible_C3
indelible_B3
indelible_A3

Figure 4. Hierarchical Clustering from datasets. Each color represents one dataset. Green palettes are *INDELible* lineages. A) Clustal control dendrogram and B) using MS-SSIM distances.

found in DNA sequence nucleotides, especially in coding sequences where close nucleotides are translated 3 by 3 into amino acids that will interact to create proteins.

Considering that the Globins datasets consist of coding sequences of proteins, it is important to note that in this study, the algorithms analyzed the sequences at every 11 nucleotides, yielding promising results. It is believed that utilizing multiples of 3, aligning with the characteristic of protein production, could further enhance the outcomes and should be explored in future researchs.

5. Conclusions

In this work we was using HVS algorithms with default large Gaussian filter (standard deviation of 1.5) blurring the images and results already indicate that HVS algorithms can perform tasks well with the new data model. With this, it is possible to explore the use of images with lower resolution and smaller size, facilitating their storage or even thinking about a database to store the new model that already improves performance for HVS application. This not only addresses the challenges of managing large-scale genomic datasets but also offers opportunities for database researchers to explore innovative approaches in data organization and retrieval.

Even with all these results exposed here, this method open opportunities for more research goals. One big issue with our approach is that it needs improvement of algorithm complexity, parallelism and indexing, because now the methodology is using the raw data without performance improvements and takes substantially more time in comparison to traditional approaches like Clustal and Blast to process whole experiment.

Another possibility for future work is the development of better comparative forms for images generated of different sizes by the new proposed data model, given that the datasets that had the worst results were those with sequences of different sizes. Along with this, it is also possible to improve the current data model including better ways to represent evolutionary patterns and compare sequences using DNA annotations or representing frameshifts (substitutions, insertions and deletions) better than now.

In conclusion, this research contributes to the field of DNA data representation, offering valuable insights and potential applications in database analysis for researchers in bioinformatics and computational genomics. Our methodology, with its unique data model and alignment-free approach, showcases the potential for further exploration and advancements in the understanding of DNA homology and related bioinformatics tasks, such that phylogeny, sequence match and perhaps protein 3D structure prediction.

6. Acknowledgments

The authors would like to acknowledge the institutions that made this work possible. PUC-Rio's Informatics Department provided crucial infrastructure support, without which the project would not have been feasible. The individual research grant from CAPES (Brazilian Ministry of Education) and the institutional grant from CNPq (Brazilian Ministry of Science and Technology) were essential in facilitating the successful completion of the study.

References

Alhijawi, B., Awajan, A., and Fraihat, S. (2023). Survey on the Objectives of Recommender Systems: Measures, Solutions, Evaluation Methodology, and New Perspectives. *ACM Computing Surveys*, 55(5):1–93.

Alsmadi, I. and Nuser, M. (2012). String Matching Evaluation Methods for DNA Comparison. *International Journal of Advanced Science and Technology*, 47.

Bakurov, I., Buzzelli, M., Schettini, R., Castelli, M., and Vanneschi, L. (2022). Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Systems with Applications*, 189:116087.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*, 112(3):535–542.

Bilotta, M., Tradigo, G., and Veltri, P. (2019). Bioinformatics Data Models, Representation and Storage. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1-3:110–116.

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, 19(2):99.

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888.

Gibbs, A. J. and Mcintyre, G. A. (1970). The Diagram, a Method for Comparing Sequences: Its Use with Amino Acid and Nucleotide Sequences. *European Journal of Biochemistry*, 16(1):1–11.

Ginalski, K., Pas, J., Wyrwicz, L. S., von Grotthuss, M., Bujnicki, J. M., and Rychlewski, L. (2003). ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Research*, 31(13):3804–3807.

Ginalski, K., von Grotthuss, M., Grishin, N. V., and Rychlewski, L. (2004). Detecting distant homology with Meta-BASIC. *Nucleic Acids Research*, 32(suppl 2):W576–W581.

Huelsenbeck, J. P. (1995). Performance of Phylogenetic Methods in Simulation. *Systematic Biology*, 44(1):17–48.

Kania, A. and Sarapata, K. (2021). The robustness of the chaos game representation to mutations and its application in free-alignment methods. *Genomics*, 113(3):1428–1437.

Leonard, S. A., Littlejohn, T. G., and Baxevanis, A. D. (2006). Common File Formats. *Current Protocols in Bioinformatics*, 16(1):A.1B.1–A.1B.9.

Lifschitz, S., Haeusler, E. H., Catanho, M., de Miranda, A. B., Molina de Armas, E., Heine, A., Moreira, S. G., and Tristão, C. (2022). Bio-Strings: A Relational Database Data-Type for Dealing with Large Biosequences. *BioTech 2022, Vol. 11, Page 31*, 11(3):31.

Löchel, H. F. and Heider, D. (2021). Chaos game representation and its applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 19:6263.

McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server issue):W20.

Mills, L. (2014). Common File Formats. *Current Protocols in Bioinformatics*, 45(1).

Namiki, Y., Ishida, T., and Akiyama, Y. (2012). Fast DNA Sequence Clustering Based on Longest Common Subsequence. In *Communications in Computer and Information Science*, volume 304 CCIS, pages 453–460. Springer, Berlin, Heidelberg.

Plataniotis, K. N. and Venetsanopoulos, A. N. (2000). *Color Image Processing and Applications*. Digital Signal Processing. Springer Berlin Heidelberg, Berlin, Heidelberg.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Siddartha, B. K. and Ravikumar, G. K. (2019). A Novel Data Masking Method for Securing Medical Image. *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*, pages 30–34.

Siddartha, B. K. and Ravikumar, G. K. (2020). An efficient data masking for securing medical data using DNA encoding and chaotic system. *International Journal of Electrical and Computer Engineering (IJECE)*, 10(6):6008.

Sievers, F. and Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science : A Publication of the Protein Society*, 27(1):135.

Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multi-scale structural similarity for image quality assessment. In *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402.