ador/doc

XIII Argentine Congress of Bioinformatics and Computational Biology

XIII International Conference of the Iberoamerican Society of Bioinformatics

III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

Unsupervised Clustering and Convolutional Networks for Learning Structures in Bioinformatics

Ignacio Fucksmann, Leandro A. Bugnon y Diego H. Milone
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), UNL-CONICET

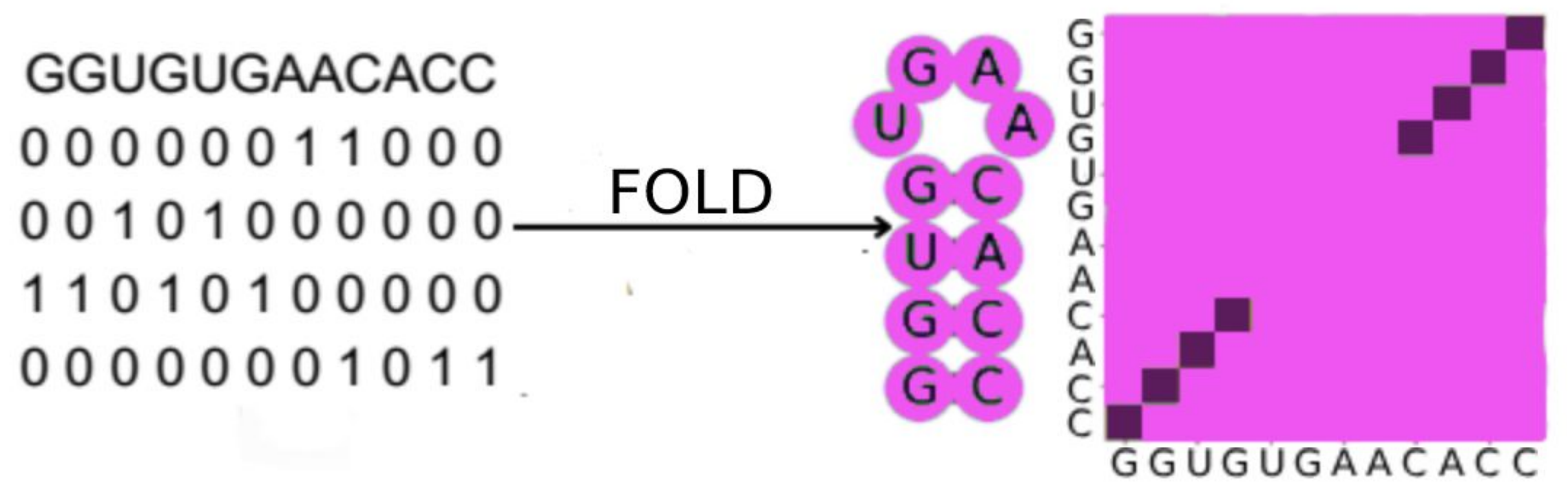
INTRODUCTION

The ability to predict structures from sequences is an important tool in bioinformatics.
In recent years, there has been an increase in the use of machine learning-based methodologies that perform at levels comparable to classical methods based on thermodynamic principles.
In this work, we propose a new model to enhance the prediction of these structures by combining unsupervised learning and deep neural networks.

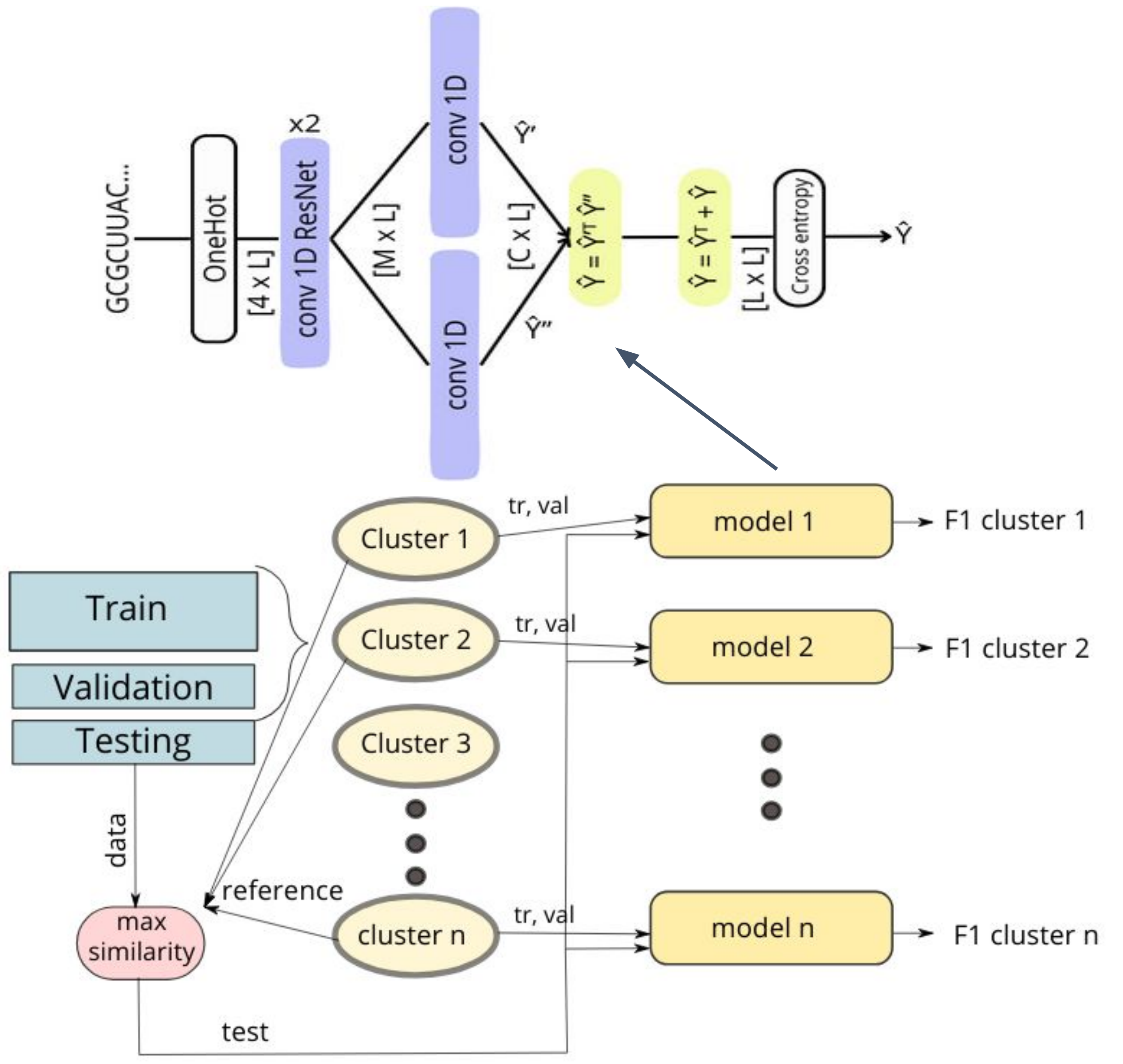
METHODS

RNA sequences are composed of four possible elements identified by the symbols A, C, G, and U, which can be numerically represented as a 4 dimension binary vector. Thus, a sequence of length L is represented as a matrix of Lx4.

In the following example, the encoding for each element and a simple example (L=12) of the encoding for the 2D structure that the model has to predict in its output are observe



The base model for prediction consists of one-dimensional (1D) convolutional layers and the subsequent encoding of LxL that relates the elements of the sequence.



The original proposal of this work involves **clustering the sequences** beforehand in an unsupervised manner and **training independent models for each group**.

For the model development:

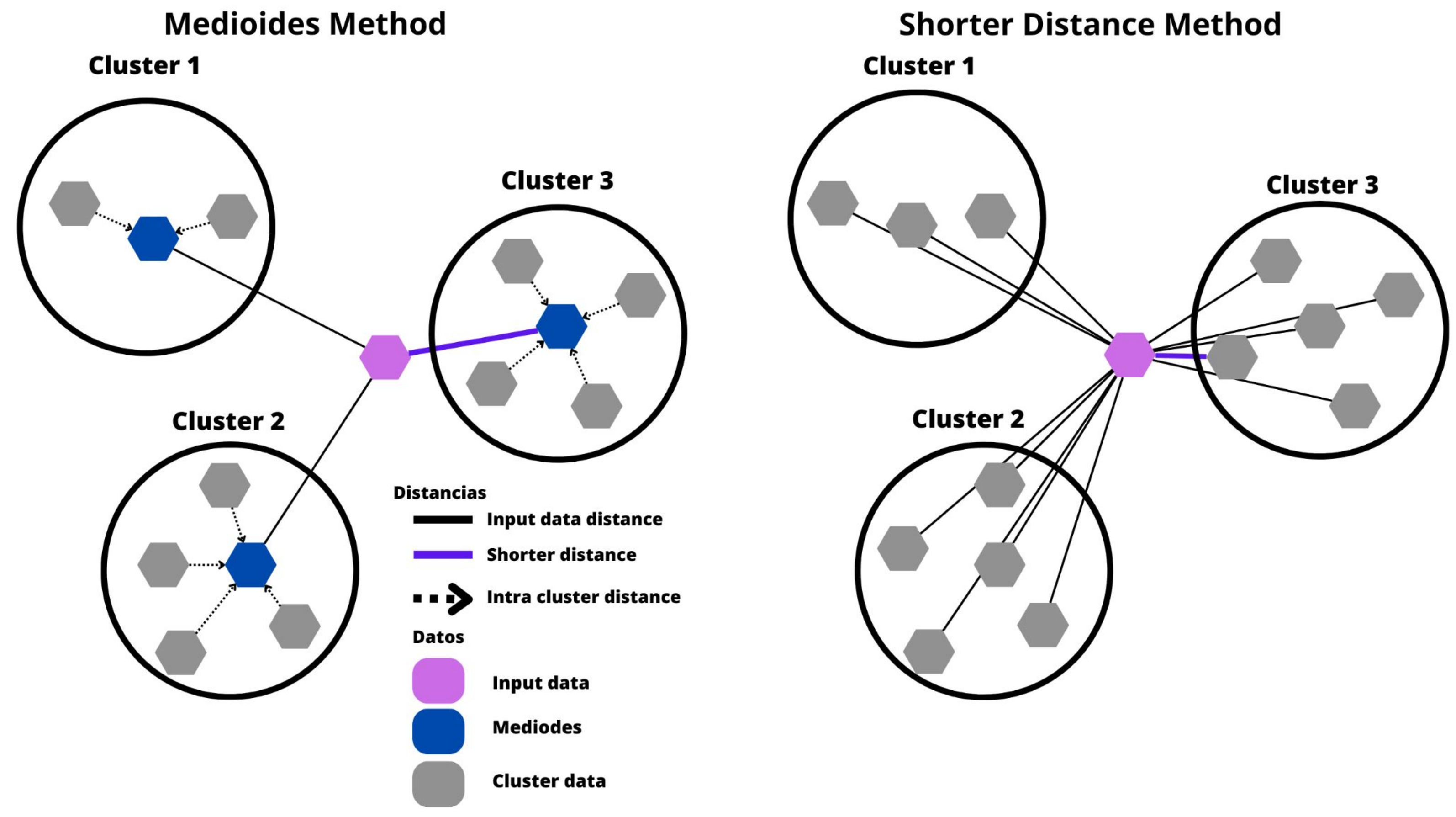
- Different clusters were generated, varying the number of groups in a range of k ∈ [10,30].
- Normalized mutual information was used to measure correspondence between the clusters based on the distance between sequences and structures.

$$NMI(Y,C) = \frac{2\ I(Y;C)}{[H(Y)+H(C)]}$$

- With the optimal k value, the data was divided into groups to train each submodel.

To predict a sequence, it is necessary to identify which group it belongs to. Two methodologies were evaluated:

- Medoid Selection: It involves identifying the sequence that is closest to all the others within its cluster. Then, the distance to each medoid is measured, and the sequence is assigned to the closest one.
- Selection by Minimum Distance: This method involves calculating the distance to each training sequence and assigning it to the group of the closest one.



RESULTS

Dataset

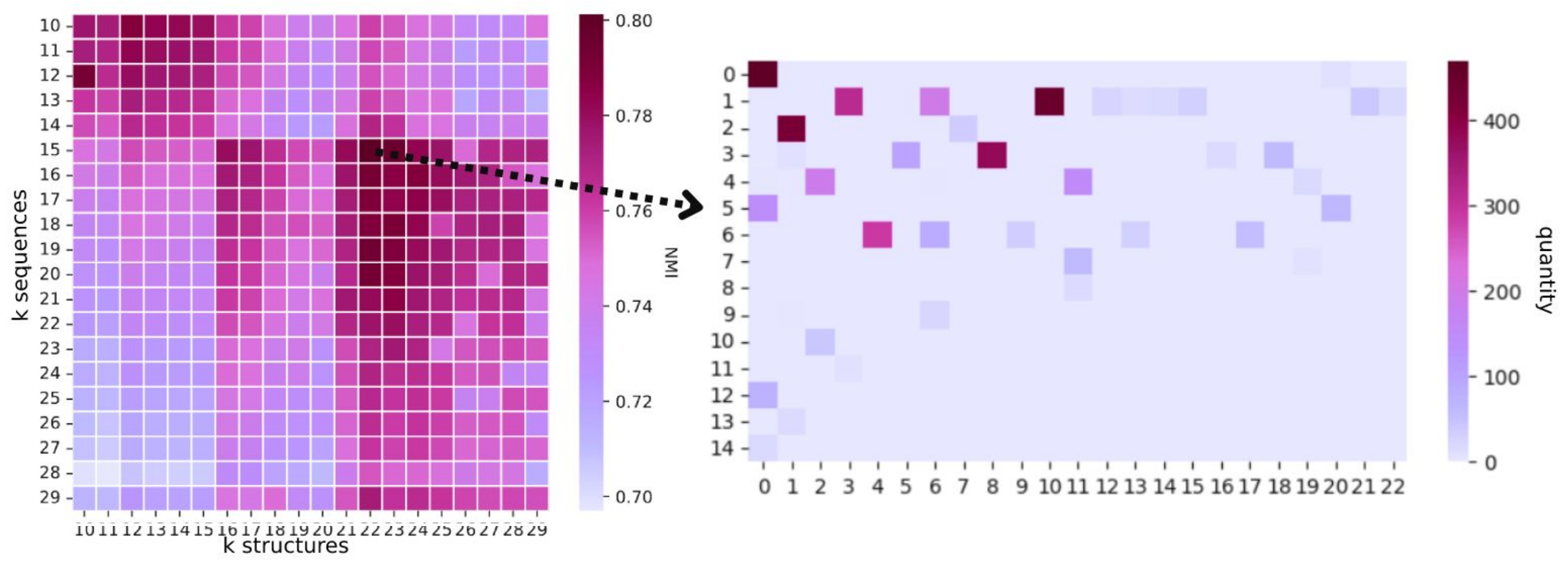
The tests were conducted using the ArchiveII dataset, which is a standard in RNA secondary structure prediction:

- 3974 sequences from various families and lengths.
- Training, validation, and test partitions with 80%, 10%, 10% respectively.

Experiments

Since at the time of prediction the model does not have access to the structures, it is necessary to determine how many clusters k based on sequence distance provide groups most similar to those obtained by structure distance.

An NMI of 0.79 (maximum) was obtained for k=22 and k=15 between the groups generated by structure and sequence distances, respectively. Therefore, in the following experiments, 15 groups obtained through unsupervised clustering based on sequence distances were used.



Once the submodels were trained, we proceeded to verify the effectiveness of the prediction by comparing the proposed model using both selection methods against the base model. The proposed model using the medoid selection method achieved an F1 score of 0.590, which did not surpass that of the baseline model (F1 = 0.769). However, when using the minimum distance selection, the results of the baseline model improved, as shown in the following table.

	Modelo base	Modelo Propuesto
Cluster 1	0.580	0.587
Cluster 2	0.890	0.956
Cluster 3	0.779	0.854
Cluster 4	0.921	0.951
Cluster 5	0.877	0.958
Cluster 6	0.567	0.623
Promedio	0.769	0.821

CONCLUSIONS

The process of unsupervised clustering and subsequent model ensemble significantly improves prediction results. On average, the proposed model achieves an F1 score that exceeds the baseline model by 5.20%. As future work, efforts will be made to enhance performance in groups with a limited number of sequences. One of the proposed approaches in this direction is to perform transfer learning from models trained on the larger groups.

REFERENCES

Bugnon L. A., Persia L. D., Gerard A. M., Edera J., Raad S., Prochetto E., Fenoy G. S., y Milone D. H. 2023. "sincFold: end-to-end learning of short-and long-range interactions for RNA folding"