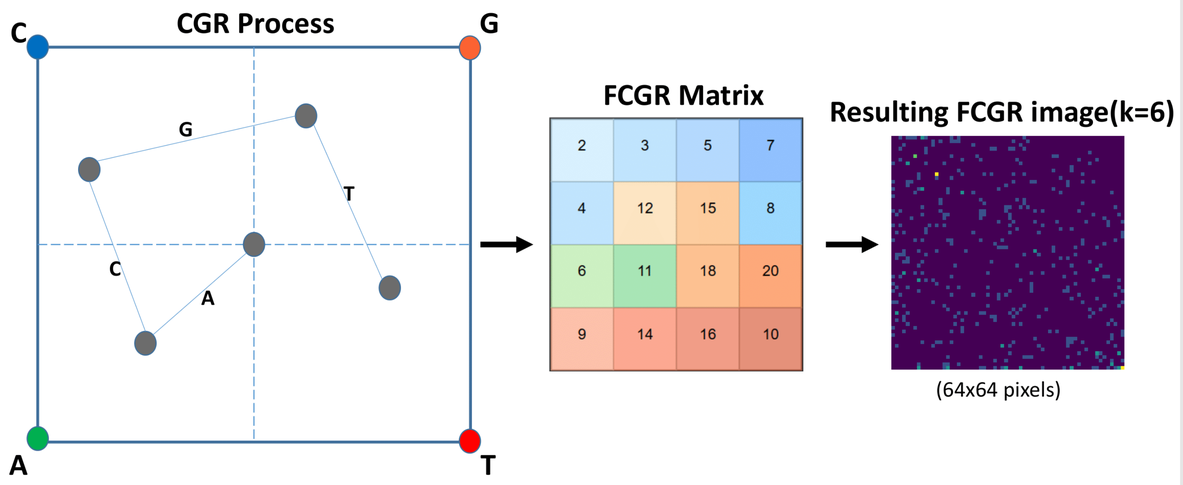
CENTER FOR BIOSYSTEMS AND BIOTECH DATA SCIENCE
Espoir Kabanga, Arnout Van Messem, Wesley De Neve

A DEEP LEARNING APPROACH TO IMAGING DINUCLEOTIDE PATTERNS AT DNA SPLICE JUNCTIONS

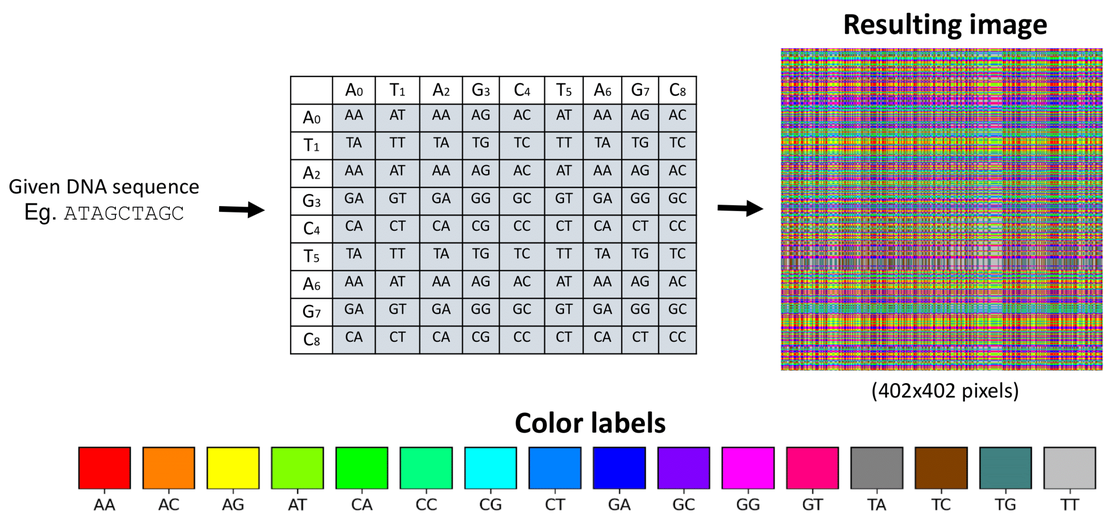
Abstract

This work introduces a novel fixed color pattern technique for visualizing DNA dinucleotides as images, enabling effective classification through deep learning models. We compare our approach against the existing Frequency Chaos Game Representation (FCGR) using ResNet50 architectures across diverse genomic datasets. Our innovative dinucleotide representation shows significantly higher performance in splice site prediction compared to FCGR and offers new perspectives for interpreting DNA sequences through image representation. We employ Grad-CAM and saliency mapping to explore model decision-making, seeking insights into the biological patterns captured by these visualization techniques.

DNA to Image Conversion



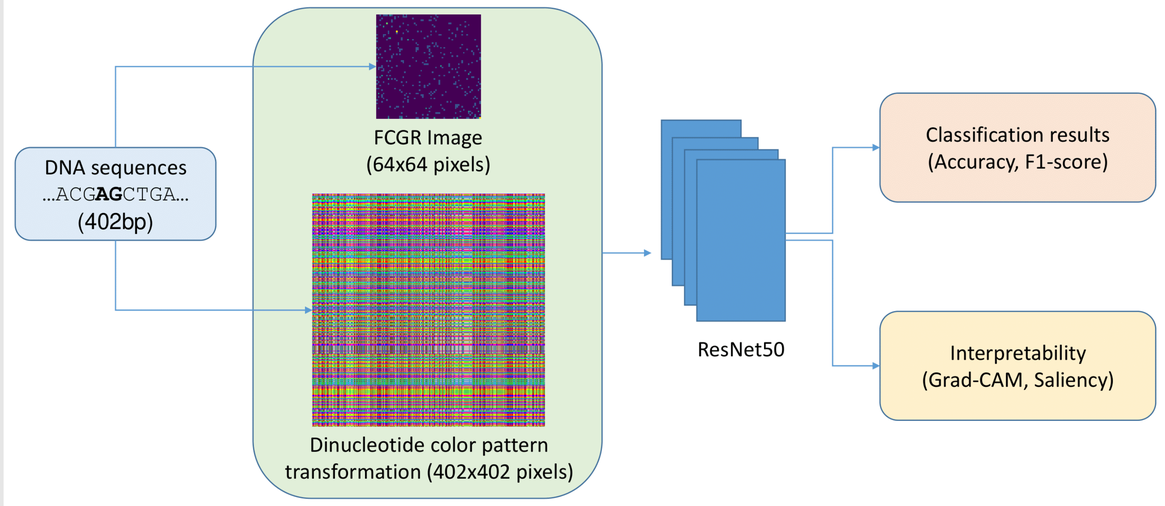
FCGR Method:
- Maps nucleotide to corners of square
- Creates chaos game representation (CGR) trajectory
- Counts frequency in grid cells
- Visualizes with intensity color maps



	A_0	T_1	A_2	G_3	C_4	T_5	A_6	G_7	C_8
A_0	AA	AT	AA	AG	AC	AT	AA	AG	AC
T_1	TA	TT	TA	TG	TC	TT	TA	TG	TC
A_2	AA	AT	AA	AG	AC	AT	AA	AG	AC
G_3	GA	GT	GA	GG	GC	GT	GA	GG	GC
C_4	CA	CT	CA	CG	CC	CT	CA	CT	CC
T_5	TA	TT	TA	TG	TC	TT	TA	TG	TC
A_6	AA	AT	AA	AG	AC	AT	AA	AG	AC
G_7	GA	GT	GA	GG	GC	GT	GA	GG	GC
C_8	CA	CT	CA	CG	CC	CT	CA	CT	CC

Dinucleotide color pattern method:
- Maps each dinucleotide to a fixed RGB color
- Creates a grid showing nucleotide interactions across sequence
- Colors each cell based on its specific dinucleotide pair
- Preserves sequence composition in a consistent image format

Methodology



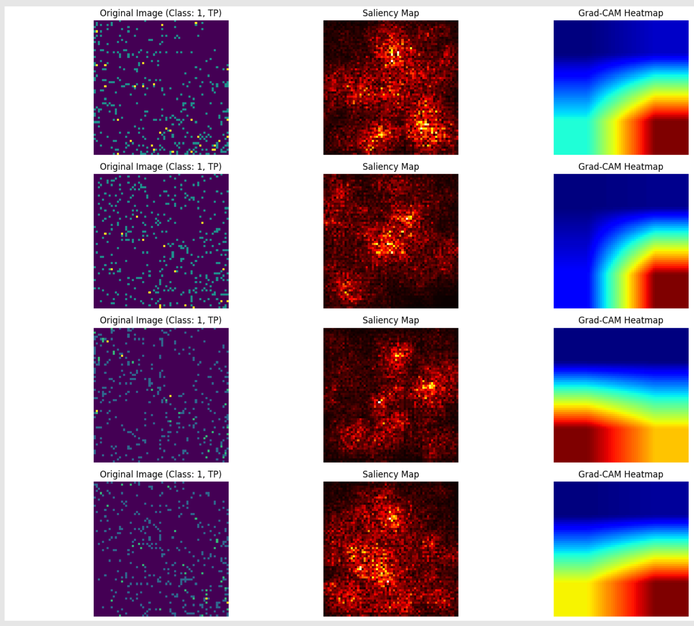
Dataset:
- *Arabidopsis thaliana* and *homo sapiens* DNA sequence
- For each species, there is a donor set (positive and negative) and acceptor set (positive and negative)
- Each DNA sequence is length 402bp
- Splice sites are located at fixed position (middle of the sequence)

Experimental setup:
- ResNet50 model used
- 50,000 images for training, 10,000 for validation and 10,000 for testing

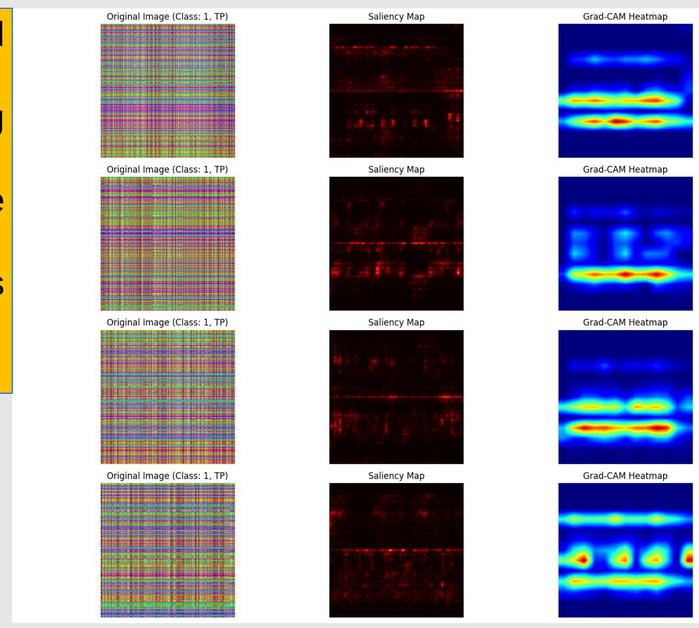
Quantitative results

Data species	Type	Image representation	Accuracy	F1-score
Arabidopsis thaliana	Donor	FCGR	0.775	0.7841
		Dinucleotide color	**0.9294**	**0.9267**
	Acceptor	FCGR	0.7723	0.7826
		Dinucleotide color	**0.9186**	**0.9158**
Homo sapiens	Donor	FCGR	0.7767	0.7852
		Dinucleotide color	**0.9573**	**0.9567**
	Acceptor	FCGR	0.7484	0.7529
		Dinucleotide color	**0.9437**	**0.9443**

Interpretability



- Dinucleotide color pattern images reveal structured information enabling precise model attention
- Model focuses on specific horizontal bands, identifying position-specific motifs
- FCGR images show diffuse attention while Dinucleotide color pattern images enable localized feature detection
- Consistent band patterns suggest the model captures biologically relevant positional information



Conclusions and future work

- Dinucleotide color pattern method significantly outperforms FCGR
- Dinucleotide color pattern method shows better interpretability
- Need to correlate model attention regions with known biological motifs to validate biological relevance

Contact: espoir.kabanga@ghent.ac.kr

GHENT UNIVERSITY GLOBAL CAMPUS
Center for Biosystems & Biotech Data Science
FACULTY OF ENGINEERING AND ARCHITECTURE
LIÈGE université Mathématique