DANCE: Deep Learning-Assisted Analysis of ProteiN Sequences Using Chaos Enhanced Kaleidoscopic Images

Authors
Taslim Murad, Prakash Chourasia, Sarwan Ali and Murray Patterson

Presented By

Murray Patterson

ICCABS 2026 North Carolina State University
February 16, 2026

Table of Contents

Introduction

- Cancer is a complex disease characterized by uncontrolled cell growth, which requires the identification of the type for effective treatment strategies.
- T cell receptors (TCRs) are essential immune proteins that recognize antigens, including cancer-associated ones.
- Advances in sequencing technologies have enabled comprehensive profiling of TCR repertoires, identifying TCRs with potent anti-cancer activity for TCR-based immunotherapies.

Motivation

- Efficient representation of TCR protein sequences is crucial for capturing their structural and functional information.
- Challenges in analyzing TCR sequences include their shorter length compared to other biomolecules and potential information loss in traditional vector-based embeddings.
- Image-based representation offers an efficient alternative, preserving critical details for comprehensive TCR sequence analysis.

Research Aim

To develop and evaluate a novel approach for analyzing T-cell receptor (TCR) protein sequences by generating image-based representations using the Chaos Game Representation (CGR) and a kaleidoscopic imaging method (DANCE).

- Preserve the structural and functional details of TCR sequences, overcoming limitations of traditional vector-based embeddings.
- Leverage these image-based representations to classify TCR protein sequences in terms of their target cancer cells.
- Explore the relationship between the visual patterns in kaleidoscopic images and the underlying biochemical properties of TCRs.
- Demonstrate the potential of combining CGR-based visualization with deep learning models for insights into TCR behavior, ultimately contributing to cancer immunotherapy advancements.

Underlying ML tasks

- To assess classification we employ metrics average accuracy, precision, recall, weighted F_1, macro F_1, ROC-AUC, and training run-time.

Chaos Game Representation (CGR)

- Converts linear amino acid sequences into 2D images.
- Maps each amino acid to unique x- and y-axis coordinates using predefined rules.
- Produces images where the spatial pixel distribution represents sequence order and amino acid composition.
- Finally we get Kaleidoscopic Images: Visually captivating 2D representations of protein sequences. Demonstrates compositional and sequential characteristics.

Methodology - 2 Steps Process

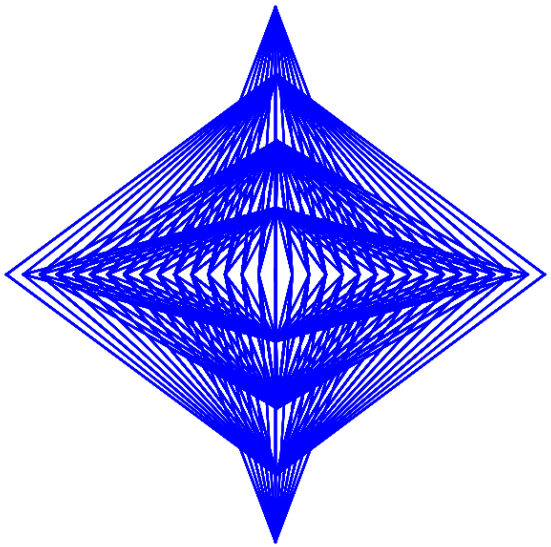
Step 1 : Assign Numerical Coordinates to Amino Acids

- Assign unique x- and y-axis values to the 20 amino acids.Assign unique x- and y-axis values to the 20 amino acids.
- Ensures distinct representation and avoids overlap in the CGR image.

Step 2: Recursively Generate DANCE Images

- Input protein sequence, recursion depth, central seed position, angle of rotation, and scale factor.
- Recursively plots coordinates, updates values, and reduces depth until termination criteria are met.
- Outputs a kaleidoscopic (symmetrical) image for the given sequence.

A sample Image Generated



Amino Acid	x-axis	y-axis	Amino Acid	x-axis	y-axis
A	0.5	0.5	M	0.5	0.0
C	1.0	0.5	N	0.25	0.5
D	0.5	1.0	P	1.0	0.0
E	0.0	0.5	Q	0.0	1.0
F	1.0	1.0	R	0.5	0.25
G	0.25	0.25	S	0.75	0.5
H	0.75	0.25	T	0.5	0.75
I	0.75	0.75	V	0.0	0.0
K	0.25	0.75	W	1.0	0.25
L	0.75	0.0	Y	1.0	0.75

- A kaleidoscopic shape image generated using chaos game representation for a sample sequence “ACQRSTAGTACGT”.
- Amino acids with corresponding x- and y-axis values.

A sample Image Generated With Depth 1

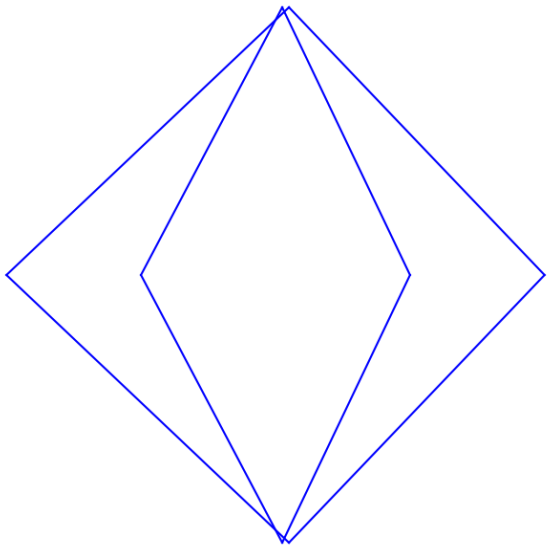


Figure: For sequence “AC”.

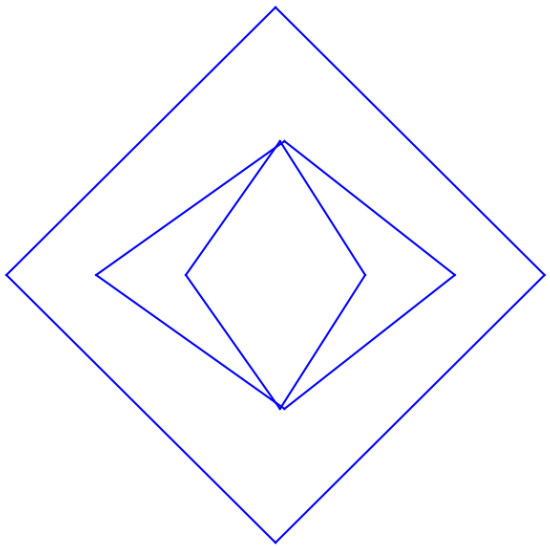


Figure: For sequence “ACQ”.

- A kaleidoscopic shape image generated using chaos game representation for a sample sequence with depth 1.

Algorithm

Algorithm Generate Kaleidoscope (DANCE)

Input: Set $\mathcal{M}$ of (m-mer) minimizers on alphabet Σ
Output: ViralVectors based embedding V

function GenKaleidoscope(seq, $depth$, pos, $angle$, $scale$)
  if $depth \leq 0$ **then**
    return
  $x, y \leftarrow pos$
  $dx \leftarrow scale \cdot \cos(angle)$
  $dy \leftarrow scale \cdot \sin(angle)$
  for *AminoAcid* **in** seq **do**
    $x, y \leftarrow x + dx, y + dy$
    $cx, cy \leftarrow$ CoordinateRule(*AminoAcid*) ▷ from Table 9
    plt.plot($[x, cx]$, $[y, cy]$, color=$color$)
    plt.plot($[x, cx]$, $[y, -cy]$, color=$color$)
    plt.plot($[-x, cx]$, $[-y, cy]$, color=$color$)
    plt.plot($[-x, cx]$, $[-y, -cy]$, color=$color$)
    GenKaleidoscope(seq, $depth - 1$, (x, y), $angle$, $scale$)
    GenKaleidoscope(seq, $depth - 1$, $(x, -y)$, $angle$, $scale$)
    GenKaleidoscope(seq, $depth - 1$, $(-x, y)$, $angle$, $scale$)
    GenKaleidoscope(seq, $depth - 1$, $(-x, -y)$, $angle$, $scale$)
    $depth \leftarrow depth - 1$

Overall Contribution

- **Introducing the use of CGR for generating kaleidoscopic images of protein sequences:** Generates visually captivating kaleidoscopic shape images that capture the structural and functional characteristics of proteins.
- **Demonstrating the effectiveness of DANCE images for protein sequence classification:** Employing deep learning image classifiers in DANCE images and demonstrating their efficacy in classifying protein sequences based on the visual patterns.
- **Investigating the relationship between visual patterns in DANCE images and protein properties:** We analyze how the kaleidoscopic shape reflects structural motifs, protein domains, secondary structures, and other relevant features.
- **Bridging the gap between visual representations and protein classification:** Addresses the gap in deep learning techniques for protein sequence classification.

Dataset Distribution

- The TCR sequence data was obtained from TCRdb [1].
- Dataset consisting of 14205 TCR sequences for four different types of cancers.

Target Label (Cancer Type)	$\lvert Sequences \rvert$
HeadNeck	5230
Ovarian	583
Pancreatic	2887
Retroperitoneal	5505

Results - TCR Classification

DL Model	Method	Acc. ↑	Prec. ↑	Recall ↑	F1 (Weig.) ↑	F1 (Macro) ↑	ROC AUC ↑	Train Time (hrs.) ↓
-	Efficient Kernel [2]	0.386	0.149	0.386	0.215	0.139	0.500	1.207
3-Layer Tab	OHE [3]	0.388	0.291	0.388	0.321	0.211	0.491	0.249
CNN	WDGRL [4]	0.436	0.339	0.436	0.358	0.236	0.510	**0.070**
4-Layer Tab	OHE [3]	0.371	0.286	0.371	0.288	0.192	0.489	0.330
CNN	WDGRL [4]	0.435	0.384	0.435	0.355	0.236	0.500	0.074
1-Layer	Chaos	0.343	0.330	0.343	0.335	0.246	0.498	4.983
CNN	DANCE (Ours)	**0.478**	0.440	**0.478**	0.312	0.278	**0.635**	3.099
2-Layer	Chaos	0.381	0.285	0.381	0.215	0.140	0.499	5.183
CNN	DANCE (Ours)	0.460	0.407	0.460	0.394	0.264	0.544	3.101
3-Layer	Chaos	0.379	0.143	0.379	0.208	0.137	0.500	6.156
CNN	DANCE (Ours)	**0.478**	**0.451**	**0.478**	**0.430**	**0.299**	0.559	3.186
4-Layer	Chaos	0.381	0.145	0.381	0.210	0.138	0.500	5.566
CNN	DANCE (Ours)	0.457	0.341	0.457	0.385	0.255	0.542	3.105
PreTrained	Chaos	0.379	0.143	0.379	0.208	0.137	0.489	7.600
RESNET50	DANCE (Ours)	0.459	0.343	0.459	0.393	0.261	0.501	8.152
PreTrained	Chaos	0.379	0.143	0.379	0.208	0.137	0.488	16.420
VGG-19	DANCE (Ours)	0.430	0.320	0.430	0.366	0.243	0.500	15.643

Result Discussion

- Feature-engineering-based baselines (OHE & WDGRL) perform worse than DANCE in all evaluation metrics except train runtime.
- DANCE outperforms efficient kernel methods, demonstrating better sequence pattern capture through image-based representations.
- Transforming sequences into images allows deep learning models to leverage spatial relationships and local dependencies for superior predictive performance.
- Convolutional architectures excel in recognizing complex visual structures, giving DANCE an edge over traditional vector-based or kernel methods.
- Performance improves with a 3-layer CNN model, achieving optimal accuracy, recall, and AUC-ROC scores.
- Models with more than 3 layers show reduced performance due to gradient vanishing, especially on smaller datasets.

Conclusion

- Combines Chaos Game Representation (CGR) with deep learning for protein sequence analysis.
- Addresses challenges in analyzing T-cell protein sequences by generating visually captivating kaleidoscopic images.
- Kaleidoscopic images capture essential details of protein sequences, including structural motifs and functional characteristics.
- We demonstrate the effectiveness of DANCE images for accurate protein sequence classification using deep learning models.
- Bridges the gap between visual representations and protein sequence classification. Contributes to a comprehensive and intuitive understanding of protein sequences.
- Opens new possibilities for innovative approaches in protein analysis and bioinformatics research.

Future Work

- Application to Other Biological Datasets: Evaluate DANCE on datasets such as coronavirus spike sequences and Zika virus sequences.
- Adoption of Advanced Deep Learning Models: Explore the use of models like Transformers for image-based classification tasks.

Thank You

Questions!!


S.-Y. Chen, T. Yue, Q. Lei, and A.-Y. Guo, "Tcrdb: a comprehensive database for t-cell receptor sequences with powerful search function," *Nucleic Acids Research*, vol. 49, no. D1, pp. D468–D474, 2021.


S. Ali, T. E. Ali, T. Murad, H. Mansoor, and M. Patterson, "Molecular sequence classification using efficient kernel based embedding," *Information Sciences*, vol. 679, p. 121100, 2024.


K. Kuzmin *et al.*, "Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone," *Biochemical and Biophysical Research Communications*, vol. 533, no. 3, pp. 553–558, 2020.


J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *AAAI conference on artificial intelligence*, 2018.