Weighted Chaos Game Representation For Molecular Sequence Classification

Taslim Murad, Sarwan Ali, Murray Patterson

Georgia State University
May 10, 2024

Table of Contents

Background

Sequence data analysis :

- Studies of Alterations in the protein sequence to classify and predict amino acid changes in SARS-CoV-2 are crucial in
  - Understanding the immune invasion and host-to-host transmission properties of SARS-CoV-2 and its variants
  - Identifying transmission patterns of each variant may help policymakers to prevent the rapid spread
  - May help in vaccine design and efficacy
- Unravel the mysteries of genetic info & its functional implications

Methods :

- Phylogenetic tree construction-based methods - a Traditional way to trace evolution.
- Later Machine Learning and Deep Learning played a major role

Motivation

- Improve performance and reduce computational cost.
- Insights into the evolutionary relationships between organisms, helping us understand the origins and diversity of life on Earth.
- Advancements in personalized medicine, identifying genetic variants associated with diseases and predicting patient responses to treatments.

Real World Application

- Genomic surveillance: Tracking the spread of pathogens in terms of genomic content
- Real time identification of new and rapidly emerging coronavirus variants
- Track the spread of known coronavirus variants in new municipalities, regions, countries and continents

Challenges

- For enabling ML/DL-based analysis, biological sequences need to be transformed into numerical representations.
- But usually the numerical feature embedding generation methods undergo sparsity and curse of dimensionality challenges.
- State-of-the-Art DL classifiers perform suboptimal on tabular data compared to tree-based methods due to their interpretability, robustness, efficiency, and feature handling capabilities..

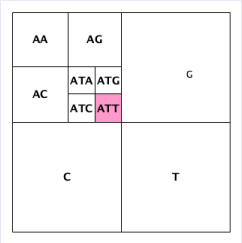
Problems

- Variable lengths of sequences
- Capturing both local and global structures
- Traditional methods (e.g. Phylogenetic Trees) are computationally expensive
- Mutations happen disproportionally

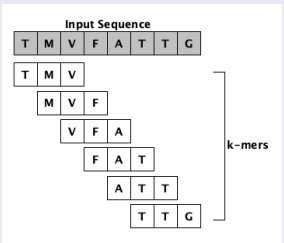
Methodology

- We propose Chaos Game Representation-based method, which is an efficient way to convert sequences into images.
- Our proposed embedding method is alignment-free and could improve the "area of interest" within the image by performing biologically meaningful manipulation of a sequence first and then mapping the manipulated sequence into an image

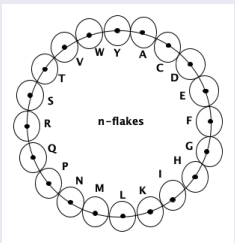
Chaos Game Representation (CGR)



(a) CGR-based allocation.



(b) 3-mers for a protein sequence



(c) 20-flakes for protein sequence.

(a) illustrates the CGR-based space allocation for a given k-mer in the respective image.(b) shows an example of 3-mers from a given sequence. (c) shows an example of 20-flakes for protein sequences.

Chaos Game Representation (CGR)

- CGR is used to convert sequences into images. Works well for nucleotide sequences.
- FCGR follows CGR to get images of protein sequences.
  - Get the x and y axis for an amino acid i using the given equations:

$$x[i] = r \cdot sin(\frac{2\pi i}{n} + \theta) \tag{1}$$

Here, r is a scaling factor that determines the size of the image, i is the position of the amino acid in the sequence, n is the total number of amino acids in the sequence, and θ is an angle parameter that affects the orientation of the image.

$$y[i] = r \cdot cos(\frac{2\pi i}{n} + \theta) \tag{2}$$

- These equations create a positional mapping of amino acids in a protein sequence onto a 2D plane, allowing the visualization of protein sequences as images. The values of r and θ can be adjusted to modify the appearance and characteristics of the resulting images.

Chaos Game Representation (CGR)

- The use of sine (sin) and cosine (cos) functions is based on trigonometry and geometric principles.
- Circular Movement: The sin and cos functions are commonly used in circular motion or circular patterns.
- Angle Variation: The angle inside the sin and cos functions $(\frac{2\pi i}{n} + \theta)$ controls the variation of positions along the circular pattern. Here: $\frac{2\pi i}{n}$ divides the circle into n equal parts based on the position of the amino acid i in the sequence. θ introduces an additional angle parameter that can rotate or shift the circular pattern, allowing for variations in the resulting image orientation.

Chaos Game Representation (CGR)

- Spatial Distribution: By combining sin and cos with the angle parameters, the equations generate a spatial distribution of points that covers the 2D space effectively. The use of trigonometric functions helps distribute the points evenly along the circular or spiral path, ensuring a balanced representation of the sequence.
- Scaling and Orientation: The scaling factor r in front of sin and cos determines the size of the circular pattern or spiral. A larger r value results in a larger pattern, while a smaller r value creates a tighter and more condensed pattern. The angle parameter θ allows for the adjustment of the image's orientation. By changing θ, we can rotate or shift the circular/spiral pattern, providing flexibility in the visual representation of the sequence.

Molecular Properties (Weights)

- Kyte and Doolittle (KD) Hydropathy Scale
  - Assigns numerical values to amino acids based on their hydrophobicity/hydrophilicity, used in predicting protein structure and function.
- Eisenberg Hydrophobicity Scale
  - Quantifies the hydrophobicity of amino acids, aiding in protein structure prediction and understanding protein interactions with hydrophobic environments.
- Hydrophilicity Scale
  - Measures the propensity of amino acids to interact with water, crucial for understanding protein solubility, folding, and function in aqueous environments.
- Flexibility Of The Characters
  - Evaluates the flexibility or rigidity of amino acids, important for predicting protein dynamics, conformational changes, and flexibility in molecular interactions.
- Hydropathy Scale
  - Ranks amino acids based on their hydrophobic or hydrophilic nature, assisting in studying protein folding, membrane protein structure, and transmembrane domains.

Workflow

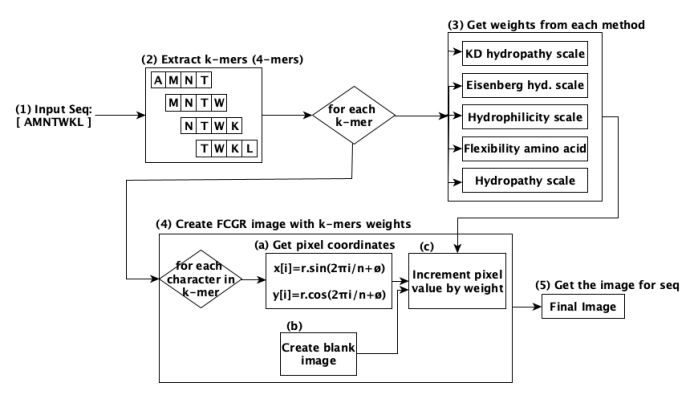


Figure: Workflow of the proposed method for creating an image of a sequence.

Classification Models

- Two types of classification models are used:
  - Tabular Models: 3-layer Tab CNN & 4-layer Tab CNN
  - Vision Models: ViT, CNN and EfficientNetB0 (pre-trained).

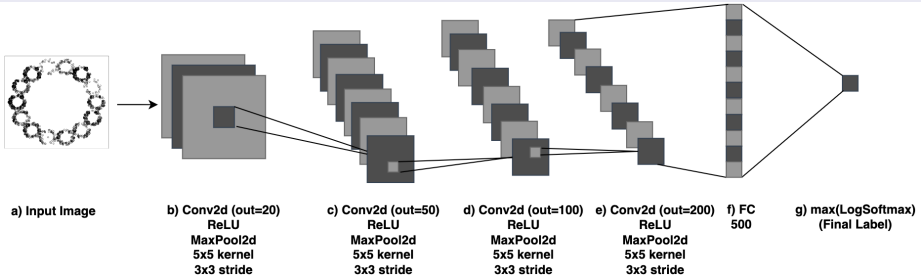


Figure: The architecture of the CNN model with 4 hidden convolution layers and softmax classification layer.

Dataset

Host Name	Count	Rabies Sequence Length			Number of Sequences		
		Min.	Max.	Average	Training	Validation	Testing
Canis Familiaris	9065	90	11928	1600.50	5802	1450	1813
Bos Taurus	2497	117	11928	995.29	1599	399	499
Vulpes Vulpes	2221	133	11930	2923.77	1422	355	444
Felis Catus	1125	90	11928	1634.43	720	180	225
Procyon Lotor	884	291	11926	6763.80	567	141	176
Desmodus Rotundus	875	164	11923	1051.50	560	140	175
Mephitis Mephitis	864	220	11929	1266.59	554	138	172
Homo Sapiens	838	101	11928	1537.85	537	134	167
Eptesicus Fuscus	718	264	11924	1144.35	460	115	143
Skunk	492	211	11928	6183.26	316	78	98
Tadarida Brasiliensis	270	264	11923	1175.67	173	43	54
Equus Caballus	202	163	11924	1376.74	130	32	40
Total	20051	-	-	-	-	-	-

Table: Dataset Statistics for Rabies data.

Baselines

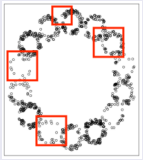
- Feature-engineering-based methods
  - One Hot Encoding (OHE): created embeddings are sparse and face curse of dimensionality challenge.
  - Wasserstein Distance Guided Representation Learning (WDGRL): require large training data for optimal performance.
  - Position Specific Scoring Matrix (PSSM)
- Image-based method
  - Frequency Matrix-based Chaos Game Representation (FCGR): 1-to-1 mapping between the amino acids and pixels.

Results

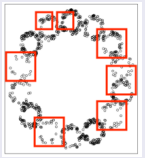
	Method	Acc. ↑	Prec. ↑	Recall ↑	F1 (Weig.) ↑	F1 (Macro) ↑	ROC AUC ↑	Train Time (Sec.) ↓
NB	OHE	0.124	0.447	0.124	0.134	0.195	0.585	979.44
	WDGRL	0.514	0.441	0.514	0.410	0.184	0.575	**0.01**
	PSSM2Vec	0.125	0.296	0.125	0.072	0.105	0.58	**0.04**
3 Layer Tab CNN	OHE	0.451	0.203	0.451	0.280	0.050	0.500	4191.34
	WDGRL	0.450	0.202	0.450	0.279	0.049	0.500	1737.65
	PSSM2Vec	0.452	0.204	0.452	0.281	0.051	0.500	2040.81
4 Layer Tab CNN	OHE	0.452	0.204	0.452	0.281	0.051	0.500	5974.26
	WDGRL	0.535	0.318	0.535	0.395	0.103	0.500	964.97
	PSSM2Vec	0.450	0.204	0.450	0.282	0.052	0.500	3790.09
ViT	Chaos	0.448	0.201	0.448	0.277	0.051	0.500	2943.45
	KD	0.440	0.194	0.440	0.269	0.050	0.500	3593.00
	Eisen.	0.465	0.216	0.465	0.295	0.052	0.500	3474.12
	Flex.	0.441	0.194	0.441	0.270	0.051	0.500	3035.72
	Hydrophil.	0.455	0.207	0.455	0.285	0.052	0.500	2829.95
	Hydropathy	0.449	0.201	0.449	0.278	0.051	0.500	3029.90
CNN	Chaos	0.780	0.763	0.780	0.767	**0.662**	**0.813**	12505.91
	KD	0.771	0.757	0.771	0.756	0.647	0.807	13331.11
	Eisen.	**0.787**	**0.779**	**0.787**	**0.773**	**0.668**	0.810	14127.47
	Flex.	0.775	0.763	0.775	0.758	0.647	0.807	13068.88
	Hydrophil.	**0.785**	**0.770**	**0.785**	**0.774**	0.659	**0.817**	14286.38
	Hydropathy	0.773	0.766	0.773	0.765	0.653	0.809	13115.00
Pretrain	Chaos	0.202	0.365	0.202	0.230	0.081	0.500	146831.05
	KD	0.210	0.370	0.210	0.229	0.079	0.510	147221.45
	Eisen.	0.284	0.451	0.284	0.364	0.095	0.530	161828.01
	Flex.	0.274	0.441	0.274	0.387	0.087	0.500	144477.50
	Hydrophil.	0.283	0.431	0.283	0.363	0.093	0.521	150921.41
	Hydropathy	0.252	0.331	0.252	0.323	0.073	0.500	142441.85

Table: The top 2 best values for each evaluation metric are shown in bold.

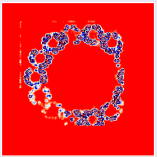
Results



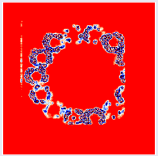
(a) Chaos



(b) Eisenberg



(c) S.M. Chaos



(d) S.M. Eisenberg

Figure: Images generated using Chaos and Eisenberg encoding techniques for a sequence against Cytoplasm location from protein subcellular dataset along with their respective Saliency Maps (S.M.). Some of the major differences between the original images are indicated using the red boxes. The blue color in the saliency maps indicates the most importance. This figure is best seen in colors.

Results

(a) Chaos

(b) Eisenberg

Conclusion and Future Work

- We propose a method to convert bio-sequences into images using the concept of CGR.
- We assign weights to pixel values in the images to enhance performance.

Future Work

- Try on larger data to evaluate the scalability.
- Employ other methods like spaced minimizers to get the images.

Thank You