Seq-HyGAN: Sequence Classification via Hypergraph Attention Network

(Khaled Mohammed Saifuddin, Corey May, Farhan Tanvir, Muhammad Ifte Khairul Islam, Esra Akbas, 2023)

Presenting to: Prof. Esra Akbas, Georgia State University
Presenter: Eljan Mahammadli, George Washington University

Eljan Mahammadli | 14.05.2024

Agenda

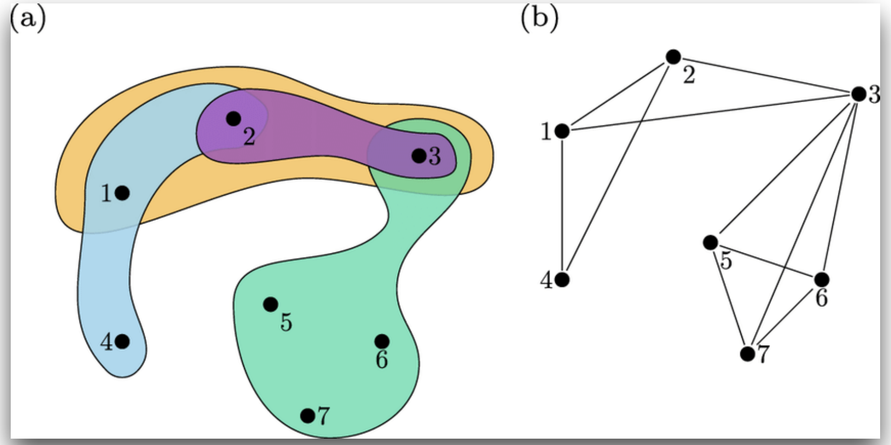
1. Introduction
2. What is a Hypergraph?
3. Seq-HyGAN Model Overview
4. Hypergraph Construction
5. Attention Mechanisms in Seq-HyGAN
6. Experimental Results
7. Practical Applications
8. Limitations and Future Work
9. Conclusion and Q&A

Introduction

- In the domain of data science, sequence classification tasks are critical for understanding complex data structures in fields like bioinformatics, text processing, and more.
- Traditional methods like RNNs and simple graph-based models often struggle with non-adjacent and complex relationships within sequence data, limiting their effectiveness.
- Graph use cases:
  - DNA-protein binding prediction
  - protein function prediction
  - drug-drug interaction prediction
- *"Approach is built on the assumption that sequences sharing structural similarities tend to belong to the same classes, and sequences can be considered similar if they contain similar subsequences."*
  - Represent sequences in a hypergraph framework, where the sequences are hyperedges that connect their subsequences as nodes.
  - Using this representation as Seq-HyGAN architecture that employs a three-level attention-based neural network

What is a Hypergraph?

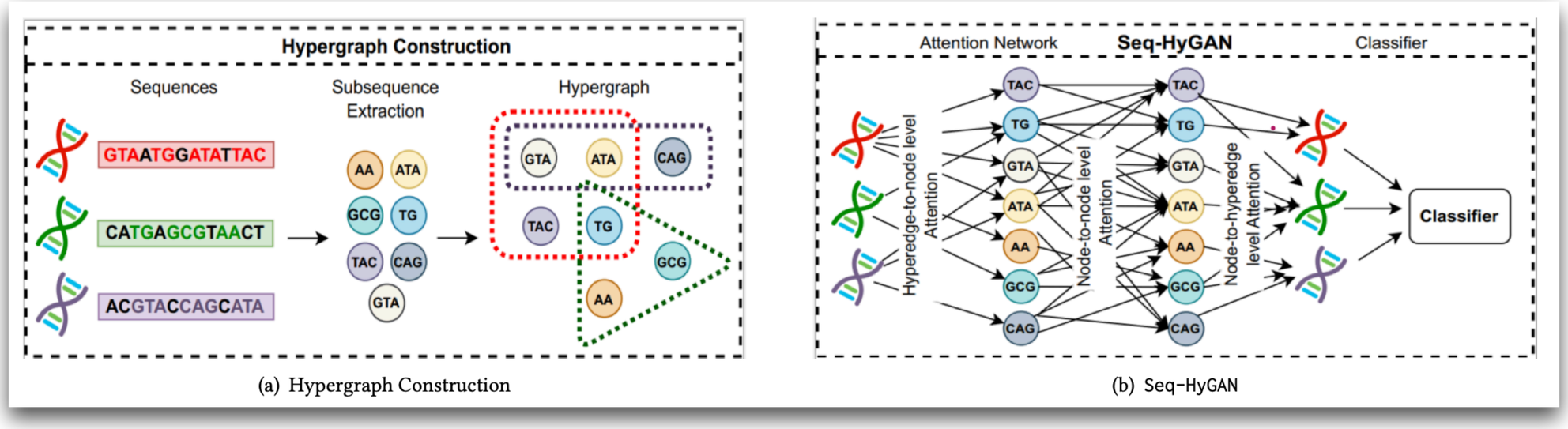
- A hypergraph extends the concept of a traditional graph by allowing edges, called **hyperedges**, to connect any number of nodes (degree-free), unlike graphs which connect only two.
- This allows for modeling complex interactions that cannot be captured by standard graphs.



Example of a hypergraph (a) and traditional graph (b)
Image credits

Seq-HyGAN Model Overview

- Seq-HyGAN introduces a novel hypergraph-based model with a three-tier attention mechanism designed to enhance sequence classification.
- This model captures both **local** and **global** structural relationships through its unique architecture.

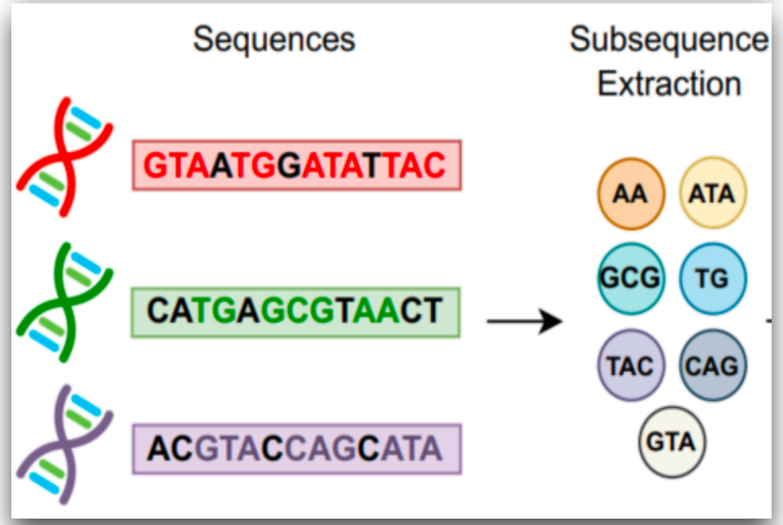


Hypergraph Construction

- In Seq-HyGAN, subsequences of sequences are treated as nodes, and sequences themselves become hyperedges that connect these nodes based on structural similarities.
- This approach allows the model to capture higher-order relationships more effectively.
- Algorithms to generate the subsequences:
  - **ESPF:** only selects the most frequent subsequences
    - we may seldom lose some infrequent important ones
  - **k-mer:** uses all the extracted subsequences for a certain k value
    - delivers better performances

Algorithm 1: Sequence Hypergraph Construction

Input: *Sequences*
Output: Hypergraph incident matrix: H
Subsequence_list<- *Sequence_Decomposition*(*Sequences*);
/* Sequence_Decomposition() could be ESPF or *k*-mer that decomposes sequences into moderated size subsequences. */
for *each subsequence in Subsequence_list* **do**
  if *subsequence is in Sequence_dictionary[sequence]* **then**
    $H[i, j] = 1$; /* i,j is the id of subsequence and sequence, respectively. */
  end
end
Output: Hypergraph incident matrix, H



Attention Mechanisms in Seq-HyGAN

Attention Mechanism

- **Hyperedge-to-Node Attention:** Focuses on the global context by aggregating information from multiple sequences to enhance the node's (subsequence's) representation.
    - representation of nodes to capture the **global context** in the hypergraph
- **Node-to-Node Attention:** Emphasizes local context by focusing on how different subsequences within the same sequence interact.
    - **local information** of nodes specific to hyperedges
    - also incorporates a position encoder that assigns a unique position to each subsequence
- **Node-to-Hyperedge Attention:** Compiles a comprehensive representation of the sequence by aggregating information back from nodes to hyperedges.
    - contribution of nodes in hyperedge construction may not be the same
    - during the aggregation process, it considers the representations of the nodes from both local and global contexts

Linear Projection

- The output from the attention layer is linearly projected to weight matrix to generate C dimensional output for each hyperedge as:

$$Z = nW_c^T$$

$$p_i^l = \mathrm{AG}_{\mathrm{E-V}}{}^l(p_i^{l-1}, n_j^{l-1} | \forall e_j \in E_i), \quad (1)$$

$$m_{i,j}^l = \mathrm{AG}_{\mathrm{V-V}}{}^l(p_i^l, p_y^l | \forall v_y \in e_j), \quad (2)$$

$$n_j^l = \mathrm{AG}_{\mathrm{V-E}}{}^l(n_j^{l-1}, p_i^l, m_{i,j}^l | \forall v_i \in e_j) \quad (3)$$

Experimental Results

- Seq-HyGAN was tested against several baseline models across multiple datasets and consistently outperformed traditional methods.
- This demonstrates the model's superior capability in handling complex sequence classification tasks.
- **Note:** The impact of the ESPF frequency threshold on Human DNA dataset.
    - a change in the frequency threshold from 5 to 25 leads to a nearly 25% reduction in the F1 score

Model	Method	Human DNA			Bach choral			Anticancer pept.		
		P	R	F1	P	R	F1	P	R	F1
ML	LR	92.82	90.64	90.84	88.09	76.68	78.52	77.00	83.16	77.67
	SVM	90.09	85.39	85.83	89.30	80.27	82.08	83.10	86.32	83.27
	DT	92.87	80.37	83.68	86.49	70.85	73.92	78.28	85.32	81.35
DL	RCNN	68.84	37.90	27.86	76.83	71.30	68.23	69.62	80.00	73.34
	BiLSTM	77.80	39.27	35.18	73.93	69.96	66.32	65.70	81.05	72.57
Node2vec	LR	36.09	32.19	22.89	16.11	20.17	18.14	71.32	82.11	75.11
	SVM	10.32	30.82	14.52	14.14	20.18	16.17	66.39	81.05	72.99
	DT	18.04	18.26	18.13	23.03	22.87	22.89	68.75	64.21	66.40
Graph2vec	LR	21.07	26.94	23.63	23.34	19.73	17.81	66.39	81.05	72.99
	SVM	10.32	30.82	14.52	13.09	15.75	16.50	71.32	82.11	75.11
	DT	19.41	19.63	19.39	25.60	25.56	25.48	77.93	73.68	75.54
GNN	DNA-GCN	96.46	96.28	96.36	85.54	85.24	85.27	83.25	83.53	83.82
	GAT	30.06	42.14	36.01	24.75	29.19	31.12	79.23	87.44	79.67
HNN	HGNN	87.03	86.82	87.12	86.12	86.89	86.93	83.82	85.42	83.97
	HyperGAT	85.13	85.33	84.11	88.09	87.44	87.45	85.33	88.42	86.68
Seq-HyGAN	ESPF	88.77	87.89	87.78	89.93	89.72	89.88	91.98	86.75	87.65
	***k*-mer**	**98.91**	**98.88**	**98.83**	**93.78**	**93.10**	**93.18**	**93.36**	**91.72**	**92.33**

Practical Applications

- The Seq-HyGAN model offers potential applications in fields requiring detailed sequence analysis such as
  - bioinformatics
  - text processing
  - complex pattern recognition

Limitations and Future Work

- Computational Intensity (three-level attention structure): The attention mechanism, while powerful, is computationally demanding.
- Data Sparsity and Overfitting: when subsequences have low occurrence frequencies

Future Work Suggestions (from personal perspective):

- Sparsity-Aware Attention Mechanisms
- Pruning and Quantization
- Adaptive Attention Layers: dynamically adjust the focus of the model based on the complexity of the data structure
- Integrate multi-head attention

Conclusion and Q&A

- Seq-HyGAN represents a significant step forward in sequence classification, offering a robust method to capture complex data relationships.
- Advanced Hypergraph Approach
- Three-Level Attention Mechanism
- Superior Performance
- Practical Applications

Thank You!