Representation Learning for Variable-Sized Multiple Sequence Alignments

Tamara Drucks
Logic and Computation

TU Wien Informatics
Institute of Logic and Computation
Databases and Artificial Intelligence Group
Supervisor: Priv.-Doz. Dr. Nysret Musliu
Co-Supervisor: Univ.-Prof. Dr. Arndt von Haeseler
Contact: tamara.drucks@aon.at

MULTIPLE WHAT?

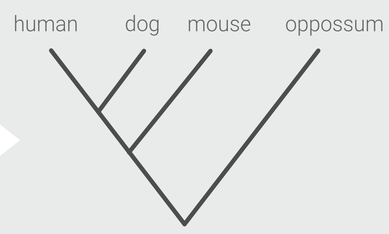
A **multiple sequence alignment**[1] refers to the alignment of three or more molecular sequences (DNA, RNA or protein), aligned such that the similarity between the sequences is maximized.

```
human     CATGGCATGTTACA...
dog       CTTCGCAACTTAAA...
mouse     GTTCGCAGGTACCA...
oppossum  CATGGCAACTTGCT...
```

Multiple sequence alignments form the basis to address numerous fundamental questions arising in biology. Many such questions stem from the research field of **phylogenetics**[1].

PHYLOGENETICS

Phylogenetics studies the evolutionary history among biological entities. The evolutionary relationships between a set of entities are typically depicted in a **phylogenetic tree**.



The reconstruction of a phylogenetic tree is based on some observable heritable traits of the given entities, such as molecular sequence data (i.e. an alignment of related sequences).

WHY LEARN REPRESENTATIONS?

Inferring phylogenies from multiple sequence alignments is hard[2–3], and remains inefficient, despite heuristics, for larger alignments. Data-driven learning approaches give hope to speed up this process. Most machine learning algorithms have some input size constraint. The computation of fixed-size representations constitutes a crucial first step.

PROBLEM STATEMENT

*Can we devise a framework that is able to produce semantically meaningful representations of **fixed size**, suitable as input for a task in phylogenetics, for **variable-sized** multiple sequence alignments?*

WHAT?

We devised a framework which has the ability to:

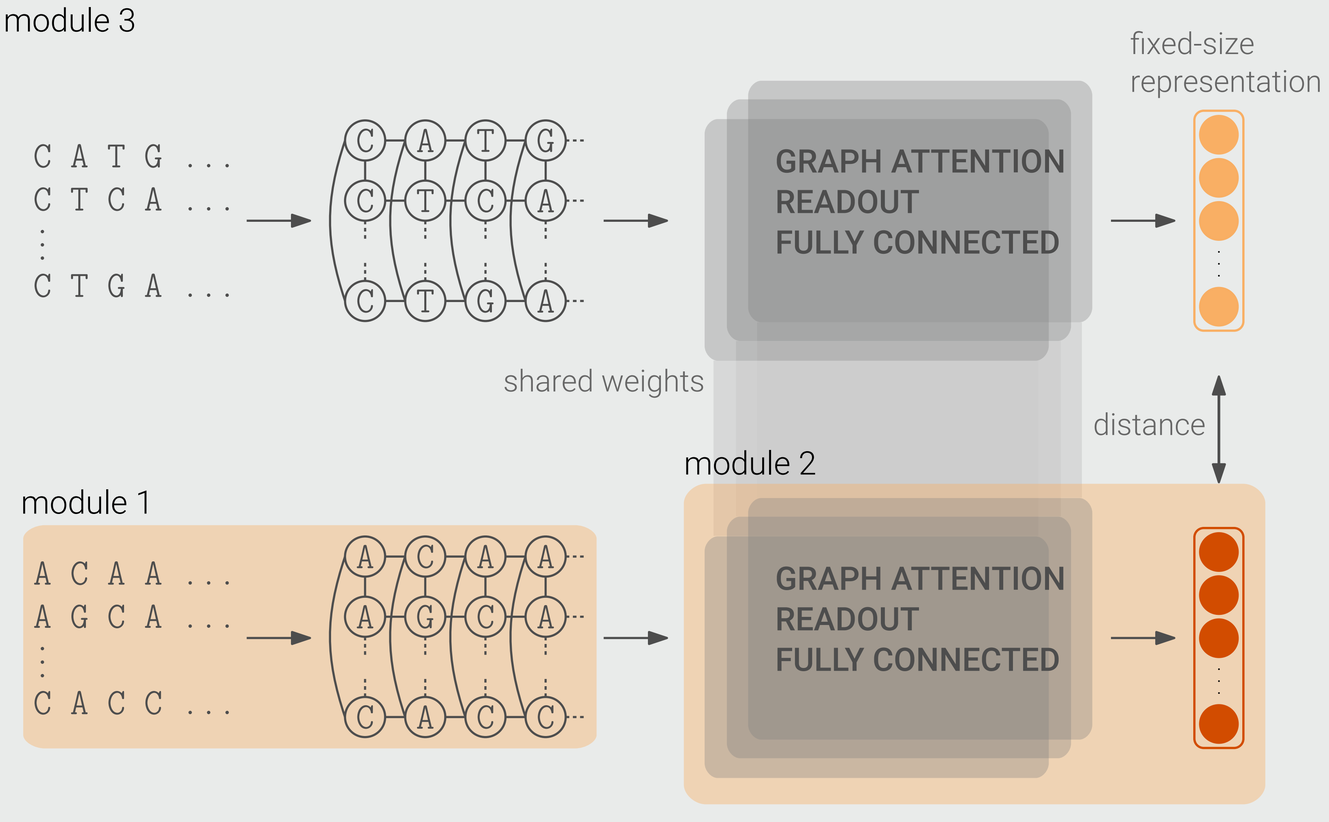
(a) Handle alignments of variable sizes

(b) Learn semantically meaningful fixed-size representations

(c) Maximize the amount of extracted information

We define semantics for alignments implicitly by means of **similarity**, i.e. similar alignments should be embedded **close** to each other, while dissimilar alignments should be embedded **distant** from each other.

REPRESENTATION LEARNING FRAMEWORK



HOW?

The framework comprises three main modules, which implement conditions **(a)** – **(c)**:

(1) Graph transformation module

(2) Embedding module

(3) Training module

(1) transforms a given alignment into a graph. (2) computes a fixed-size representation using graph attention layers with average pooling as graph readout, followed by fully-connected layers. (3) implements the training procedure using a siamese neural network with contrastive loss function. Training is done with pairs.

EXPERIMENTAL EVALUATION & BENCHMARK STUDY

We conducted systematic experiments to assess the representation learning framework. The phylogenetic task of the selection of the **model of sequence evolution**[1] **(mSE)** serves as learning objective for a proof of concept. We simulated alignments evolving under up to four different mSEs (JC, K2P, F81 and GTR).

WHAT DID WE LEARN?

The learned representations are semantically meaningful, given our notion of semantics.

Performance generally improved with the size of the alignment, the framework thus seems to be able to maximize the extracted information.

t-SNE of alignments evolved under two (above) and three (below) mSEs show that similar alignments are close, while dissimilar alignments are distant to each other in the embedding space.

HOW DO WE DO?

We compared our classifier *siamSE* with the established methods[1] for model selection AIC, AICc and BIC.

mSE	*siamSE*	AIC	AICc	BIC
JC	94.0	71.7	79.0	**98.7**
K2P	95.7	95.7	97.3	**99.7**
F81	92.3	90.0	92.0	**93.3**
GTR	96.0	**100.0**	**100.0**	98.0
mean	94.5	89.3	92.0	**97.4**

siamSE ranks second best on average with 94.5% accuracy.

IT WORKS, WHAT'S NEXT?

The results of our empirical evaluation are promising. What are possible next steps?

1. Include more complex models of sequence evolution
2. Provide more formal notions and guarantees
3. Learn representations for a different phylogenetic task, such as e.g. tree topology
4. Explore other approaches for this problem
5. Compile benchmark dataset

1. M. Steel. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing* (2010), **2.** W.H.E. Day, D.S. Johnson, and D. Sankoff. *The computational complexity of inferring rooted phylogenies by parsimony* (1986) and **3.** B. Chor and T. Tuller. *Maximum likelihood of evolutionary trees is hard* (2005).