### Representation Learning for Variable-Sized Multiple Sequence Alignments

**Tamar 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: tamar.drucks@aon.at

---

#### MULTIPLE WHAT?

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

**Example:**
- **human** CATGCGATTTACA...
- **dog** CCTCGCAACTTAAA...
- **mouse** GTTCGCAGTACCA...
- **opposum** CATGGCAACTGCT...

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

---

#### WHY LEARN REPRESENTATIONS?

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

---

#### 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 (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?**

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.

---

#### 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.

---

#### 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.

---

#### HOW DO WE DO?

We compared our classifier siamSE with the established methods$^1$ for model selection AIC, AICc and BIC.

<table>
<thead>
<tr>
<th>mSE</th>
<th>siamSE</th>
<th>AIC</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>94.0</td>
<td>71.7</td>
<td>79.0</td>
<td>98.7</td>
</tr>
<tr>
<td>K2P</td>
<td>95.7</td>
<td>95.7</td>
<td>97.3</td>
<td>99.7</td>
</tr>
<tr>
<td>F81</td>
<td>92.3</td>
<td>90.0</td>
<td>92.0</td>
<td>93.3</td>
</tr>
<tr>
<td>GTR</td>
<td>96.0</td>
<td>100.0</td>
<td>100.0</td>
<td>98.0</td>
</tr>
<tr>
<td>mean</td>
<td>94.5</td>
<td>89.3</td>
<td>92.0</td>
<td>97.4</td>
</tr>
</tbody>
</table>

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

---

**Footnotes:**


---

**References:**