Simulation of ancient DNA sequences using transformer-based techniques.

Théo Boury $^{1,2}$, Jazeps Medina-Tretmanis $^{3}$, Maria Avila-Arcos $^{4}$, Emilia Huerta-Sanchez $^{3}$, Burak Yelmen $^{2,5}$, Flora Jay $^{2}$

$^{1}$Computer Science Department, Ecole Normale Supérieure de Lyon, France; $^{2}$U Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, France; $^{3}$Center for Computational Molecular Biology, Brown U, USA; $^{4}$International Laboratory for Human Genome Research, U Nacional Autónoma de México, México; $^{5}$Institute of Genomics, U of Tartu, Estonia.
Ancient DNA specificities

Undamaged DNA

\[
\text{A T C G T}
\]

Damaged Ancient DNA

\[
\text{A ? A C A A ? C G A}
\]

Difficulties with Ancient DNA (aDNA):

- Degrades over time
- Contaminated by external DNA
- More missing data and errors than modern DNA
Why study Damaged Ancient DNA?

Undamaged DNA

Damaged Ancient DNA

A ? A C A A ? C G A

Ancestor tree

-3.8 billions years

-1 Million year

Present day
DNA sequences for inference purposes....
Inference tool

Ancient DNA sequences for training
Sparse and highly damaged
No labels
AT?GCT???GCGTATT AT??ATCCT??A????????
TAT????TCTCC?C?GT

Ancient DNA sequences for test
ATCGCTC??????T?A
AT????TGCTCCAGT

Training

Testing

aDNA sequences for inference purposes....
... required simulation of aDNA sequences

- Inference tool
- Training
- Testing

Simulated ancient DNA for training
Numerous with custom parameters
Generated with labels

| AT??CTCTCC??GT   | ATTA????TGCT??CAG |
| ATCGCGTACTC???? | ATG??CTC????CCAGT |
| AT?GCT???GCATT   | AT??ATCCT??A??????? |
| TAT????TCTCC?C?GT | ???????ATCGCTC?????? |

Ancient DNA sequences for test

| ATCGCTC????????T?A  | ATC????????CCCAG |
State-of-the-art aDNA simulator: Gargammel

Gargammel complexity: \( O(n \times c \times f) \)

With \( c \), the desired coverage and \( f \), the number of fragments sampled by Gargammel

\( f \) can lead to a large overhead in practice

---

1 Renaud et al, 2016, Bioinformatics
Achieved result: our new seq-to-seq aDNA simulator

- Generate aDNA sequences from undamaged ones
- Method: Iterative process over a specific encoder-only transformer
- Data simulation: Undamaged sequences: Msprime$^2$. Damaged sequences: pipeline around Gargammel$^3$
- Generation complexity: $O(n^3)$, of interests compared to Gargammel

$^2$Baumdicker et al, 2021, Genetics
$^3$Jazeps et al, 2023, ?
Attention interest\textsuperscript{4} versus convolution

\textbf{Convolution}

- Slow weights update
- $O(n)$ complexity

\textbf{Attention}

- Quick weights update
- $O(n^2)$ complexity

\textsuperscript{4}Vaswani et al, 2017
Our use of the pretrained DNABERT model\textsuperscript{5}

Max seq. length: 512 nucleotides. Comp.: 12 Transformer layers

\textsuperscript{5} Ji et al, 2021, Bioinformatics
Our data and our fine-tunings

A mask prediction task to predict the aDNA part

A binary classification task to measure if the aDNA part is a plausible “translation” of the undamaged DNA
Generation algorithm using Mask Prediction only

- Complexity with Mask Prediction alone: $O(n^3)$
- With $n$, size of the input

Inspired by bert-gen, Wang et al, 2019
Use of classification as a complement for mask prediction in the generation

Classification: Is the aDNA sequence a “plausible” translation for the undamaged sequence?

Complexity with the $K$-Top table: $O(Kn^3)$
Results

- We do alignments\(^7\) with 30 chunks of aDNA
- Aligned sequences are identical at 74 percent in average
- We count similar positions to the exclusion of gaps and missing data

\(^7\)Wheeler et al, 2000, Nucleic Acids Research
Perspective: push further complexity and performances

- **First leverage**: Sparse\(^8\) or linear\(^9\) attention instead of full attention
  Reduce attention to \(O(n)\) instead of \(O(n^2)\).

- **Second leverage**: Use of SNPs
  Counteracts the 512 nucleotide limitation and "diminishes" \(n\)

<table>
<thead>
<tr>
<th>Full sequences</th>
<th>Intermediates</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T T A G G A C G</td>
<td>A A G G</td>
<td>0 1 1 0</td>
</tr>
<tr>
<td>A T T A C G A C A</td>
<td>A A C A</td>
<td>0 1 0 1</td>
</tr>
<tr>
<td>C T T C G G A C G</td>
<td>C C G G</td>
<td>1 0 1 0</td>
</tr>
</tbody>
</table>

Positions 1 2 3 4 5 6 7 8 9

---

\(^8\)Zaheer et al, 2021
\(^9\)Nesterenko et al, 2022
Conclusion and future work

▶ A new seq-to-seq simulation technique for ancient DNA sequences
▶ Complexity in $O(n^3)$ in simpler case, complemented to the use of batches in practice

Future work:

▶ Define new criteria to assess the quality of sequences
▶ Do our own pre-training instead of using DNABERT’s one.
▶ Define an ”in-between” classification task when using a K-Top table
▶ General advance in transformers will be at our advantage:
  Integration of sparse attention
Acknowledgements

Contributors:

- Jazeps Medina-Tretmanis
- Maria Avila-Arcos
- Emilia Huerta-Sanchez
- Antoine Szatkownik (‘‘Generative model for genomics’’ poster!)
- Burak Yelmen
- Flora Jay
- Guillaume Charpiat
- YOU!
Our ancient DNA simulation using encoder-transformer