A Manuscript Microscope 'Sentence Audit' of:

Accel-Align: A Fast Sequence Mapper and Aligner Based on the Seed–Embed–Extend Method

What is the Manuscript Microscope Sentence Audit?

The Manuscript Microscope Sentence Audit is a research paper introspection system that parses the text of your manuscript into minimal sentence components for faster, more accurate, enhanced proofreading.

Why use a Sentence Audit to proofread your manuscript?

- Accelerated Proofreading: Examine long technical texts in a fraction of the usual time.
- Superior Proofreading: Detect subtle errors that are invisible to traditional methods.
- Focused Proofreading: Inspect each individual sentence component in isolation.
- Reliable Proofreading: Ensure every single word of your manuscript is correct.
- Easier Proofreading: Take the hardship out of crafting academic papers.

Bonus 1: Improved Productivity: Rapidly refine rough drafts to polished papers.
Bonus 2: Improved Authorship: Cultivate a clear, concise, consistent, writing style.
Bonus 3: Improved Reputation: Become known for rigorously precise publications.

Manuscript Source: https://www.biorxiv.org/content/10.1101/2020.07.20.211888v2

Manuscript Authors: Yiqing Yan, Nimisha Chaturvedi & Raja Appuswamy

Audit Date: 31/03/21 **Audit Identifier:** TN09BNWFC7U19H0 **Code Version:** 3.6

Features of the Sentence Audit:

The Sentence Audit combines two complementary proofreading approaches:

1. Each sentence of your text is parsed and displayed in isolation for focused inspection.
2. Each individual sentence is further parsed into Minimal Sentence Components for a deeper review of the clarity, composition and consistency of the language you used.

The Minimal Sentence Components shown are the smallest coherent elements of each sentence of your text as derived from it's conjunctions, prepositions and selected punctuation symbols (i.e. commas, semicolons, round and square brackets).

The combined approaches ensure easier, faster, more effective proofreading.

Comments and Caveats:

• The sentence parsing is achieved using a prototype natural language processing pipeline written in Python and may include occasional errors in sentence segmentation.

• Depending on the source of the input text, the Sentence Audit may contain occasional html artefacts that are parsed as sentences (E.g. "Download figure. Open in new tab").

• Always consult the original research paper as the true reference source for the text.

Contact Information:

To get a Manuscript Microscope Sentence Audit of any other research paper, simply forward any copy of the text to John.James@OxfordResearchServices.com.

All queries, feedback or suggestions are also very welcome.

Research Paper Sections:

The sections of the research paper input text parsed in this audit.

Section No.	Headings	Sentences
Section: 1	Abstract	16
Section: 2	1 Introduction	11
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0

The Sentence Audit Of The Research Paper

Title **Accel-Align: A Fast Sequence Mapper and Aligner Based on the Seed–Embed–Extend Method**

S1 [001] Abstract

S1 [002] Background

Background

S1 [003] Improvements in sequencing technology continue to drive sequencing cost towards $100 per genome.

Improvements ...
... in sequencing technology continue ...
... to drive sequencing cost towards $100 ...
... per genome.

S1 [004] However, mapping sequenced data to a reference genome remains a computationally-intensive task due to the dependence on edit distance for dealing with indels and mismatches introduced by sequencing.

However, ...
... mapping sequenced data ...
... to a reference genome remains a computationally-intensive task ...
... due to the dependence ...
... on edit distance ...
... for dealing ...
... with indels ...
... and mismatches introduced ...
... by sequencing.

S1 [005] All modern aligners use seed–filter–extend (SFE) methodology and rely on filtration heuristics to reduce the overhead of edit distance computation.

All modern aligners use seed–filter–extend ...
... (SFE) ...
... methodology ...
... and rely ...
... on filtration heuristics ...
... to reduce the overhead ...
... of edit distance computation.

S1 [006] However, filtering has inherent performance–accuracy trade-offs that limits its effectiveness.

However, ...
... filtering has inherent performance–accuracy trade-offs ...
... that limits its effectiveness.

S1 [007] Results

Results

S1 [008] Motivated by algorithmic advances in randomized low-distortion embedding, we introduce seed– embed–extend (SEE), a new methodology for developing sequence mappers and aligners.

Motivated ...
... by algorithmic advances ...
... in randomized low-distortion embedding, ...
... we introduce seed– embed–extend ...
... (SEE), ...
... a new methodology ...
... for developing sequence mappers ...
... and aligners.

S1 [009] While SFE focuses on eliminating sub-optimal candidates, SEE focuses instead on identifying optimal candidates.

While SFE focuses ...
... on eliminating sub-optimal candidates, ...
... SEE focuses instead ...
... on identifying optimal candidates.

S1 [010] To do so, SEE transforms the read and reference strings from edit distance regime to the Hamming regime by embedding them using a randomized algorithm, and uses Hamming distance over the embedded set to identify optimal candidates.

To do so, ...
... SEE transforms the read ...
... and reference strings ...
... from edit distance regime ...
... to the Hamming regime ...
... by embedding them ...
... using a randomized algorithm, ...
... and uses Hamming distance ...
... over the embedded set ...
... to identify optimal candidates.

S1 [011] To show that SEE performs well in practice, we present Accel-Align, an SEE-based short-read sequence mapper and aligner that is 3-12× faster than state-of-the-art aligners on commodity CPUs, without any special-purpose hardware, while providing comparable accuracy.

To show ...
... that SEE performs well ...
... in practice, ...
... we present Accel-Align, ...
... an SEE-based short-read sequence mapper ...
... and aligner ...
... that is 3-12× faster ...
... than state-of-the-art aligners ...
... on commodity CPUs, ...

... without any special-purpose hardware, ...

... while providing comparable accuracy.

S1 [012] Conclusions

Conclusions

S1 [013] As sequencing technologies continue to increase read length while improving throughput and accuracy, we believe that randomized embeddings open up new avenues for optimization that cannot be achieved by using edit distance.

As sequencing technologies continue ...

... to increase read length ...

... while improving throughput ...

... and accuracy, ...

... we believe ...

... that randomized embeddings open up new avenues ...

... for optimization ...

... that cannot be achieved ...

... by using edit distance.

S1 [014] Thus, the techniques presented in this paper have a much broader scope as they can be used for other applications like graph alignment, multiple sequence alignment, and sequence assembly.

Thus, ...

... the techniques presented ...

... in this paper have a much broader scope ...

... as they can be used ...

... for other applications ...

... like graph alignment, ...

... multiple sequence alignment, ...

... and sequence assembly.

S1 [015] Availability

Availability

S1 [016] https://github.com/raja-appuswamy/accel-align-release

https://github.com/raja-appuswamy/accel-align-release

S2 [017] 1 Introduction

S2 [018] Over the last decade, DNA sequencing technology has achieved dramatic improvements in both cost and throughput.

Over the last decade, ...

... DNA sequencing technology has achieved dramatic improvements ...

... in both cost ...

... and throughput.

S2 [019] With the $100-per-genome sequencing goal emerging as a realistic target in the near future, the amount of genomic data generated by sequencing is only poised to grow faster.

With the $100-per-genome sequencing goal emerging ...
... as a realistic target ...
... in the near future, ...
... the amount ...
... of genomic data generated ...
... by sequencing is ...
... only poised ...
... to grow faster.

S2 [020] The first, and often one of the most time consuming steps, in analyzing genomic datasets is sequence alignment–the process of determining the location in the reference genome of each sequencing read.

The first, ...
... and often one ...
... of the most time consuming steps, ...
... in analyzing genomic datasets is sequence alignment–the process ...
... of determining the location ...
... in the reference genome ...
... of each sequencing read.

S2 [021] Sequence alignment can be boiled down to a string matching problem.

Sequence alignment can be boiled down ...
... to a string matching problem.

S2 [022] Given a string G as the reference genome, and a set of substrings R as the sequencing reads, the task of read alignment is to find the origin location of each read r ■ R in G. However, due to sequencing errors, and differences between the reference genome and the sequenced organism, a read might not exactly align at any candidate location in the reference genome.

Given a string G ...
... as the reference genome, ...
... and a set ...
... of substrings R ...
... as the sequencing reads, ...
... the task ...
... of read alignment is ...
... to find the origin location ...
... of each read r ■ R ...
... in G. However, ...
... due to sequencing errors, ...
... and differences ...
... between the reference genome ...
... and the sequenced organism, ...
... a read ...
... might not exactly align ...
... at any candidate location ...

End of Sample Audit

This is a truncated Manuscript Microscope Sample Audit.

To get the full audit of this text (or any other research paper),
forward a copy of the research paper to John James at
John.James@OxfordResearchServices.com
