OXFORD

(GIGA)n SCIENCE

GigaScience, 2017, 1–5

doi: xx.xxxx/xxxx
Manuscript in Preparation
Paper

PAPER

Smash++: finding rearrangements

Morteza Hosseini[1,*], Diogo Pratas[2,*] and Armando J. Pinho[2,*]

[1,2]IEETA/DETI, University of Aveiro, Portugal

[*]seyedmorteza@ua.pt; pratas@ua.pt; ap@ua.pt

Abstract

The Abstract (250 words maximum) should be structured to include the following details: **Background**, the context and purpose of the study; **Results**, the main findings; **Conclusions**, brief summary and potential implications. Please minimize the use of abbreviations and do not cite references in the abstract.

Key words: Keyword1; keyword 2; keyword 3 (Three to ten keywords representing the main content of the article)

Introduction to this Template

This is the epigraph text, should you like to add one. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

—Epigraph source name

This is the LaTeX template for GigaScience journal manuscript submissions. **Please note that whilst this template provides a preview of the typeset manuscript for submission, it will not necessarily be the final publication layout.**

There are important commands in the preamble that you will need to modify for your own manuscript. If you are using this template on Overleaf, please switch the editor to Source code mode to view them; or if you prefer to stay in the Rich Text view, click on the title in the Rich Text view to display the preamble code.

Use the `\journal{...}` command in the preamble so that the correct journal name, logo and colours are loaded automatically. **Only certain journals and options are supported at this time;** check with your journal's editorial office if your journal is supported.

Alternatively you can re-define `\jname`, `\jlogo` and the `jcolour` explicitly, though check with your journal's editorial office to confirm that this is appropriate.

Specify your manuscript's category with the `\papercat{...}` command in the preamble.

See the sample code in the preamble for a sample of how author and affiliation information can be specified.

Use later sections starting with 'Background' on page 2 to write your manuscript. The remainder of this current section will provide some sample LaTeX code for various elements you may want to include in your manuscript.

Sectional Headings

You can use `\section{...}`, `\subsection{...}` commands to add more sections and subsections to your manuscript. Further sectional levels are provided by `\subsubsection`, `\paragraph` and `\subparagraph`.

Citations and References

Use the `num-refs` document class option for numerical citations, and `alph-refs` option for author-year citations. Use the `\citep` command for parenthetical citations, and `\citet` command for text citations (when using `alpha-refs`). This is a citation: [1] and here are two more: [2, 3].

- This is a bullet list.
- Another point.
- A third point.

This[1] is a footnote. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et

1 This is the footnote text. This is the footnote text. This is the footnote text. This is the footnote text. This is the footnote text.

Compiled on: May 28, 2019.
Draft manuscript prepared by the author.

Key Points

- This is the first point
- This is the second point
- One last point.

dolore magna aliqua.

- This is a numbered list.
- Another point.
- A third point.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

This is a 3rd level heading

Use `\subsubsection` to get a 3rd level heading. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

This is a 4th level heading. Use `\paragraph` to get a 4th level heading. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

THIS IS A 5TH LEVEL HEADING. Use `\subparagraph` to get a 5th level heading. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Figures and Tables

Figures and tables can be added with the usual `figure` and `table` environments, e.g. Figure ?? and Table ??. Use `figure*` and `table*` if you need a two-column wide figure or table, as in Figure ?? and Table ??.

If you have a very wide table or figure, you can use `sidewaystable` or `sidewaysfigure`, as in Table ??: this will be rotated sideways and occupy a *single column* on its own.

If your table or figure is both wide and tall (so it wouldn't fit well in a single column with `sidewaystable` or `figure`), you can use `table` or `figure` inside a `landscape` environment for a full-page landscaped alternative. A page break will be inserted *immediately before and after* the `landscape` environment (Table ??), so you'll need to carefully position it in a suitable location in your manuscript.

Some Mathematics Sample

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i \qquad (1)$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

Background

The background section should be written in a way that is accessible to researchers without specialist knowledge in that area and must clearly state—and, if helpful, illustrate—the background to the research and its aims. The section should end with a brief statement of what is being reported in the article.

Data Description

A statement providing background and purpose for collection of these data should be presented for readers without specialist knowledge in that area. A brief description of the protocol for data collection, data curation and quality control, as well as potential uses should be included, as well as outlining how the data can be accessed if it is not deposited in our repository.

Analyses

This section should provide details of all of the experiments and analyses that are required to support the conclusions of the paper. The authors should make clear the goal of each analysis and state the basic findings.

Discussion

The discussion should spell out the major conclusions and interpretations of the work, including some explanation on the importance and relevance of the dataset and analysis. It should not be restatement of the analyses done and their basic conclusions. The discussion section can end with a concluding paragraph that clearly states the main conclusions of the research along with directions for future work. Summary illustrations can be included.

Potential implications

Authors should provide some additional comments about potential, more broad-ranging implications of their work that are not directly related to the current focus of their manuscript. This section is meant to promote discussion on possible ways the findings or data presented might be used in or have a relationship with other areas of research that may not be directly apparent in the work. It is not meant to provide 'proof of importance' of the work. Only to engender expansion of use to other areas.

Explicit personal opinions by the authors are permitted, but they should be made clear as such. References or related information to support the propositions should be included. These section should focus on work that can be done within the foreseeable future and specifically using the information within the

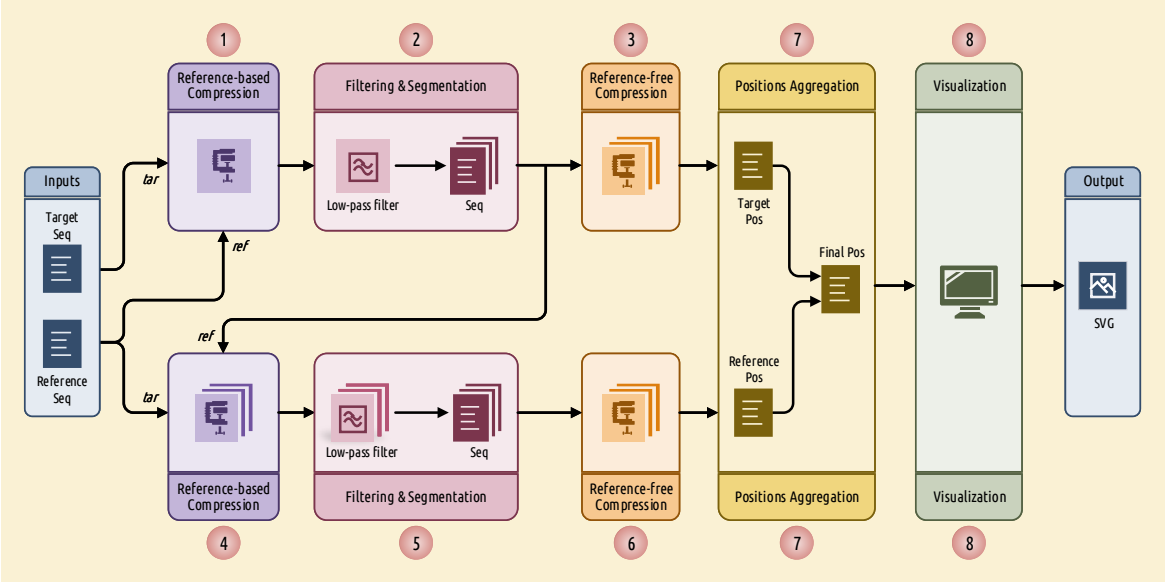

Figure 1. The schema of Smash++. The process of finding similar regions in reference and target sequences and also, computing redundancy in each region includes eight stages. Finally, Smash++ outputs a *.pos file that includes the positions of the similar regions, and can be then visualized, resulting in an SVG image.

manuscript, not provide speculation on how it will relate to far-reaching goals of the research area.

Methods

The schema of the proposed method is illustrated in Figure 1. Smash++ takes as inputs a reference and a target file and produces as output a position file, which is then fed to the Smash++ visualizer to produce an SVG image. This process has eight major stages: (1) compression of the original target file, based on the model of original reference file, (2) filtering and segmentation of the compressed file, (3) reference-free compression of the segmented files, obtained by the previous stage, (4) compression of the original reference file, based on the model of segmented files obtained by stage 2, (5) filtering and segmentation of the compressed files, (6) reference-free compression of the segmented files, that are obtained by the stage 5, (7) aggregating positions, generated by stages 3 and 6, and (8) visualizing the positions. The following sections describe the process in detail.

Data modeling

Smash++ works on the basis of cooperation between finite-context models (FCMs) and substitutional tolerant Markov models (STMMs). Applying these models on various contexts provides probability and weight values, illustrated in Figure 2a, which are then mixed (by multiplication and addition, shown in Figure 2b) to provide the final probability (P) of occurring an input symbol. The following subsections describe FCMs and STMMs in detail.

Finite-context model (FCM)

A finite-context model considers Markov property to estimate the probability of the next symbol in an information source, based on the past k symbols (a context of size k) [? 4?]. Denoting the context as $c_{k,i} = s_{i-k}s_{i-k+1}\dots s_{i-2}s_{i-1}$, the probability of the next symbol s_i in an information source S, which is posed

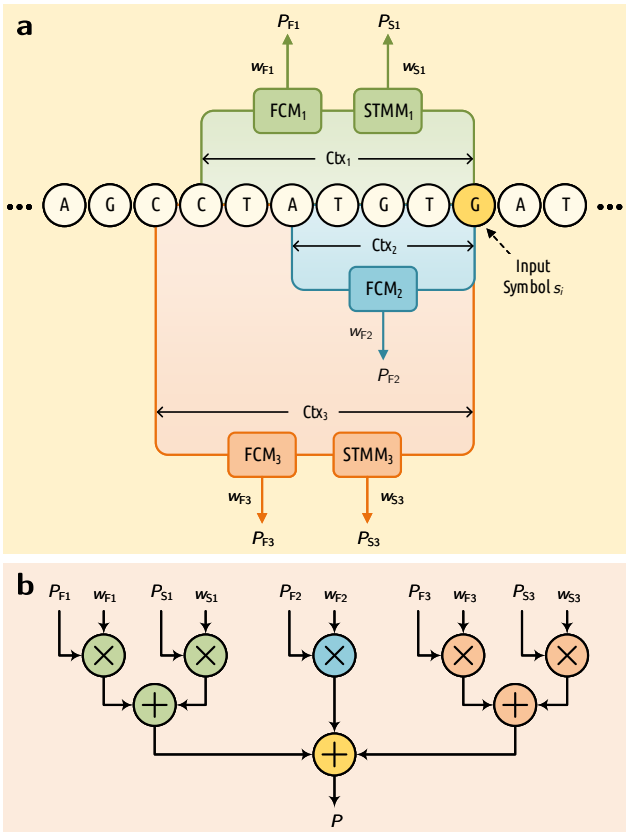

Figure 2. Data modelling by Smash++. (a) cooperation between finite-context models (FCMs) and substitutional-tolerant Markov models (STMMs). Note that each STMM needs to be associated with an FCM. (b) probability of an input symbol is estimated by employing the probability and weight values that have been obtained from processing previous symbols.

at i, can be estimated as

$$P_m(s_i|c_{k,i}) = \frac{N(s_i|c_{k,i}) + \alpha}{N(c_{k,i}) + \alpha|\Theta|}, \quad (2)$$

in which m stands for model (FCM in this case), $N(s_i|c_{k,i})$ shows the number of times that the information source has generated symbol s_i in the past, $|\Theta|$ denotes size of the alphabet Θ, $N(c_{k,i}) = \sum_{b\in\Theta} N(b|c_{k,i})$ represents the total number of events occurred for the context $c_{k,i}$ and α allows to keep a balance between the maximum likelihood estimator and the uniform distribution. Eq. 2 turns to the Laplace estimator, for $\alpha = 1$, and also behaves as a maximum likelihood estimator, for large number of events i [5].

Substitutional tolerant Markov model (STMM)

A substitutional tolerant Markov model [6] is a probabilistic-algorithmic model that assumes at each position, the next symbol in the information source is the symbol which has had the highest probability of occurrence in the past. This way, an STMM ignores the real next symbol in the source. Denoting the past k symbols as $c_{k,i} = s_{i-k}s_{i-k+1}\dots s_{i-2}s_{i-1}$, the probability of the next symbol s_i, can be estimated as

$$P_m(s_i|c'_{k,i}) = \frac{N(s_i|c'_{k,i}) + \alpha}{N(c'_{k,i}) + \alpha|\Theta|}, \tag{3}$$

where N represents the number of occurrences of symbols, that is saved in memory, and $c'_{k,i}$ is a copy of the context $c_{k,i}$ which is modified as

$$c'_{k,i} = \arg\max_{\forall b\in\Theta} P_m(b|c'_{k,i}). \tag{4}$$

STMMs can be used along with FCMs to modify the behavior of Smash++ in confronting with nucleotide substitutions in genomic sequences. These models have the potential to be disabled, to reduce the number of mathematical calculations and consequently, increase the performance of the proposed method. Such operation is automatically performed using an array of size k (the context size), named history, which preserves the past k hits/misses. Seeing a symbol in the information source, the memory is checked for the symbol with the highest number of occurrences. If they are equal, a hit is saved in the history array; otherwise, a miss is inserted into the array. Before getting to store a hit/miss in the array, it is checked for the number of misses and in the case they are more than a predefined threshold t, the STMM will be disabled and also the history array will be reset. This process is performed for each symbol in the sequence.

This example shows the distinction between a finite-context model and a substitutional tolerant Markov model. Assume, the current context at position i is $c_{11,i}$ = GGCTAACGTAC, and the number of occurrences of symbols saved in memory is A = 10, C = 12, G = 13 and T = 11. Also, the symbol to appear in the sequence is T. An FCM would consider the next context as $c_{11,i+1}$ = GCTAACGTACT, while an STMM would consider it as $c'_{11,i+1}$ = GCTAACGTACG, since the base G is the most probable symbol, based on the number of occurrences stored in memory.

Availability of source code and requirements (optional, if code is present)

Lists the following:

- Project name: e.g. My bioinformatics project
- Project home page: e.g. http://sourceforge.net/projects/mged
- Operating system(s): e.g. Platform independent
- Programming language: e.g. Java
- Other requirements: e.g. Java 1.3.1 or higher, Tomcat 4.0 or higher
- License: e.g. GNU GPL, FreeBSD etc. Any restrictions to use by non-academics: e.g. licence needed

This needs to be under an Open Source Initiative approved license where practicable compiled running software is made available. If the code is not hosted in a repository the *GigaScience* GitHub repository is also available for this purpose.

Availability of supporting data and materials

GigaScience requires authors to deposit the data set(s) supporting the results reported in submitted manuscripts in a publicly-accessible data repository such as *Giga*DB (see *Giga*DB database terms of use for complete details). This section should be included when supporting data are available and must include the name of the repository and the permanent identifier or accession number and persistent hyperlinks for the data sets (if appropriate). The following format is recommended:

"The data set(s) supporting the results of this article is(are) available in the [repository name] repository, [cite unique persistent identifier]."

Following the Joint Declaration of Data Citation Principles, where appropriate we ask that the data sets be cited where it is first mentioned in the manuscript, and included in the reference list. If a DOI has been issued to a dataset please always cite it using the DOI rather than the less stable URL the DOI resolves to (e.g. http://dx.doi.org/10.5524/100044 rather than http://gigadb.org/dataset/100044). For more see:

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/datacitation]

A list of available scientific research data repositories can be found in res3data and BioSharing.

Declarations

List of abbreviations

CPU: central processing unit; FCM: finite-context model; RAM: random access memory; STMM: substitutional-tolerant Markov model;

Ethical Approval (optional)

Manuscripts reporting studies involving human participants, human data or human tissue must:

- include a statement on ethics approval and consent (even where the need for approval was waived)
- include the name of the ethics committee that approved the study and the committee's reference number if appropriate

Studies involving animals must include a statement on ethics approval and have been treated in a humane manner in line with the ARRIVE guidelines.

See our editorial policies for more information.

If your manuscript does not report on or involve the use of any animal or human data or tissue, this section is not applicable to your submission. Please state "Not applicable" in this section.

Consent for publication

If your manuscript contains any individual person's data in any form, consent to publish must be obtained from that person, or

in the case of children, their parent or legal guardian. All presentations of case reports must have consent to publish. You can use your institutional consent form. You should not send the form to us on submission, but we may request to see a copy at any stage (including after publication). Please also confirm you have followed national guidelines on data collection and release in the place the research was carried out, for example confirming you have Ministry of Science and Technology (MOST) approval in China.

If your manuscript does not contain any individual person's data, please state "Not applicable" in this section.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by Programa Operacional Factores de Competitividade – COMPETE (FEDER); and by national funds through the Foundation for Science and Technology (FCT), in the context of the projects [UID/CEC/00127/2013, PTCD/EEI-SII/6608/2014] and the grant [PD/BD/113969/2015].

Author's Contributions

The individual contributions of authors to the manuscript should be specified in this section. Guidance and criteria for authorship can be found in our editorial policies. We would recommend you follow some kind of standardised taxonomy like the CASRAI CRediT (Contributor Roles Taxonomy).

Acknowledgements

Please acknowledge anyone who contributed towards the article who does not meet the criteria for authorship including anyone who provided professional writing services or materials.

Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements section. If you do not have anyone to acknowledge, please write "Not applicable" in this section.

See our editorial policies for a full explanation of acknowledgements and authorship criteria.

Group authorship: if you would like the names of the individual members of a collaboration group to be searchable through their individual PubMed records, please ensure that the title of the collaboration group is included on the title page and in the submission system and also include collaborating author names as the last paragraph of the "Acknowledgements" section. Please add authors in the format First Name, Middle initial(s) (optional), Last Name. You can add institution or country information for each author if you wish, but this should be consistent across all authors.

Please note that individual names may not be present in the PubMed record at the time a published article is initially included in PubMed as it takes PubMed additional time to code this information.

Authors' information (optional)

You may choose to use this section to include any relevant information about the author(s) that may aid the reader's interpretation of the article, and understand the standpoint of the author(s). This may include details about the authors' qualifications, current positions they hold at institutions or societies, or any other relevant background information. Please refer to authors using their initials. Note this section should not be used to describe any competing interests.

References

1. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Statist 2004;32:928–61.
2. Cox DR. Regression models and life tables (with Discussion). J R Statist Soc B 1972;34:187–220.
3. Heard NA, Holmes CC, Stephens DA. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. J Am Statist Assoc 2006;101:18–29.
4. Hosseini M, Pratas D, Pinho AJ. AC: A Compression Tool for Amino Acid Sequences. Interdisciplinary Sciences: Computational Life Sciences 2019;11(1):68–76.
5. Pratas D, Silva RM, Pinho AJ, Ferreira PJ. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. Scientific reports 2015;5:10203.
6. Pratas D, Hosseini M, Pinho AJ. Substitutional tolerant Markov models for relative compression of DNA sequences. In: International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) Springer; 2017. p. 265–272.