ISSN 2186-7437

NII Shonan Meeting Report

No. 197

Computational Pangenomics

Paola Bonizzoni (University of Milano-Bicocca, Italy)
Alberto Policriti (University of Udine, Italy)
Kunihiko Sadakane (The University of Tokyo, Japan)

February 19–24, 2023

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Computational Pangenomics

Organizers:
Paola Bonizzoni (University of Milano-Bicocca, Italy)
Alberto Policriti (University of Udine, Italy)
Kunihiko Sadakane (The University of Tokyo, Japan)

February 19–24, 2023

Background and introduction

Computational Pangenomics encompasses different research efforts for transitioning the existing paradigm from a sequence-based reference genome to a pan-genome, i.e., an evolutionarily coherent collection of genomes. Such a transition is urgently needed to effectively exploit the data masses produced by the technical advances and the widespread adoption of sequencing technologies. Graph-based representations of collections of genomes and diploid-aware assemblers have been recently proposed, but a large amount of work is still needed to shift to a pan-genomic view into the current research practice. Indeed, the traditional approach considers a single sequence as a reference genome, and that sequence has been obtained starting from sample tissues of unknown donors, and it has been refined through the integration of different samples. So, the human reference genome is actually the fusion of several individuals' genomes, where the characteristics of each single genome is lost. This approach led to important contributions to our understanding of human physiology and of several pathologies like cancer. However, it was essentially motivated by the limits of the early sequencing technologies and of the associated costs. In the recent years, new sequencing technologies have revolutionized the field by increasing the throughput (i.e., the amount of sequences produced in a single run), by increasing the length of the produced sequences (longer sequences allow to better disambiguate repetitive regions in the genome), by increasing the quality of the base calls (having less errors allows to reliably capture variations among individuals) while costs dramatically decreased (sequencing can be almost considered as a routine task). The resulting wealth of data bears the promise of a new course for precision medicine (i.e., adapting treatments to each individual's genetic profile). For example, thanks to the advancements of sequencing technologies, it is now possible to characterize the genetic content of a single cell and this has profound implications in the study of the evolution of cancer, where the genetic content of different cancer cells may be different due to the progressive accumulation of mutations occurred during the replication of the cancer cells. Finally, we should note that, even if the human genome was the main focus in the early days of Bioinformatics, we are assisting a spread in the use of sequencing technologies for the characterization of a growing number of species. For example, widespread sequencing efforts of the novel SARS-CoV-2

virus played a central role in the response to the pandemic, since the characterization of virus variations are aiding in tracking the international spread and in the development of the vaccines. From the computational perspective, the core problem is now how to find, to represent, and to query/compare a very large set of genetic variations obtained from large collections of genomes with the ultimate goal of making sense of such a wealth of data both for improving our understanding of the underlying biological mechanisms and for implementing the promises of translational precision medicine. Some initial and promising representations have been proposed—either based on (multi)graphs or on indexes of highly-repetitive collections of strings—but much further work is needed to really perform the transition to novel practical representations of the reference pan-genome and to novel algorithms able to exploit them.

As a consequence, the development of computational pangenomic must be sustained by coordinated efforts of different highly-specialized research areas: starting from research in stringology and indexing (for developing novel and efficient representations of the pangenome) to research in any area of Bioinformatics (for transitioning existing algorithms to the new paradigm) and to research in the data mining area (for further exploring new applications and discovering potential new associations between genetic variations and phenotypic traits). The meeting aims to provide an occasion for researchers of these research areas to present recent advances and to foster the questions that will drive the future research efforts in computational pangenomics.

Overview of the meeting

The Computational Pangenomics meeting has been attended by researchers that are experts in various topics including data structures for indexing and compressing pangenome graphs, automata theory for pangenomics, algorithms for the alignment to textual pangenome representations, computational methods for the detecting genomic variations (i.e. structural variations, SNPs) and for genome analysis in comparative pangenomics. In particular, the meeting has seen the participation of researchers from two different communities, one more involved on combinatorial pattern matching and data compression, and the other community more involved in computational methods in bioinformatics. Thus the meeting has been a good occasion for a fruitful exchange of ideas and the cross-fertilization of different computational approaches and knowledge. In particular, both communities have been working on algorithms in bioinformatics and have presented during the meeting different point of views in pangenome graph analysis.

The meeting had talked sessions where participants have presented their current work and have discussed open problems.

Overview of Talks

Structural Variation Discovery from sample-specific strings

Luca Denti, University of Milan-Bicocca, Italy

SVDSS (Structural Variation Discovery from sample-specific strings) is a

new method for discovery of SVs from PacBio HiFi reads that combines and effectively leverages mapping-free, mapping-based and assembly-based methodologies for overall superior SV discovery performance. Although the increased accuracy, there is still room for improvements and most limitations are still not fully solved yet. For instance, long read alignment in repetitive regions of the genome and SVs calling in heterozygous regions are still quite inaccurate. During the Shonan meeting, we aimed to improve long read alignment in repetitive regions of the genome by performing an ad-hoc local realignment that prefers higher consistency around potential variation over higher alignment score. To this aim, we formulated a new computational problem and we discussed different methodologies to solve it.

Algorithms for Computing Co-lex Order of Automata

Sung-Hwan Kim, University of Udine, Italy

With the recent advancement of the sequencing technologies and computational capabilities, now it is required to process a large number of reference sequences at the pangenomic scale. For indexing pangenomic graphs represented as (non)deterministic finite automata, computing the co-lexicographical order is an essential procedure. Despite of some remarkable results on particular special cases, there is still room to be improved for general cases, especially for one efficient both from a theoretical and practical point of view. This talk gives a brief survey on the state-of-the-art algorithms for computing co-lexicographical order of automata. In particular, three main techniques covering several important algorithms for deterministic and non-deterministic automata are discussed as well as the future direction of further improvement.

Pangenomic FM-indexes

Travis Gagie, Dalhousie University, Canada

DNA alignment has been a killer app for the FM-index, but aligning DNA reads against a single genome can bias research results and medical diagnoses. In the past few years we have found ways to FM-index datasets of thousands of genomes, but researchers want the results expressed in terms of compact representations called pangenome graphs. Hundreds of matches in the dataset may correspond to only one or two matches in the graph. Given a read, therefore, we would like to find which parts of it match well and where they match in the graph, in time depending on the length of the read and the number of matches in the graph but not on the number of matches in the dataset. We are now closing in on that goal; this talk will give a high-level view of the challenges and some potential solutions.

Indexing regular languages with co-lex order

Nicola Prezza, University Ca-Foscari, Italy

NFAs are inherently unordered objects, but they represent regular languages on which one can very naturally define a total order: for example, the co-lexicographic order in which words are compared alphabetically from right to

left. In this talk I will show that interesting things happen when one tries to map this total order to the states of an accepting NFA for the language: the resulting order of the states is a partial pre-order whose width p turns out to be an important parameter for NFAs and regular languages. For example, take the classic powerset determinization algorithm for converting an NFA of size n into an equivalent DFA: while a straightforward analysis shows that the size of the resulting DFA is at most 2^n, we prove that it is actually at most $(n - p + 1) * 2^p$. This implies that PSPACE-complete problems such as NFA equivalence or universality are actually easy on NFAs of small width p (the case p=1 - total order - is particularly interesting). Another implication of this theory is that we can compress NFAs to just $O(logp)$ bits per transition while supporting fast membership queries in the substring closure of the language.

Compressibility-Aware Quantum Algorithms on Strings

Sharma Valliyil Thankachan, NC State University, USA

Quantum algorithms have been established for many basic problems on strings. This work shows that new, faster quantum algorithms are possible when the string is highly compressible. We focus on two popular methods for compression – the Lempel-Ziv77 algorithm (LZ77) and the Run-length-encoded Burrows-Wheeler Transform (RL-BWT), and provide optimal quantum algorithms. We also show an efficient way of constructing a (known) compact index with equivalent capabilities as the suffix tree. This data structure is then applied to numerous problems, such as the longest common substring, finding maximal unique matches, lyndon factorization, etc (see arXiv:2302.07235 for a preliminary version).

Analyzing SARS-CoV-2 waste water samples, by Deconstructing a pangenome

Tomas Vinar, Comenius University, Slovakia

The genomes of SARS-CoV-2 are classified into variants, some of which are monitored as variants of concern (e.g. the Delta variant B.1.617.2 or Omicron variant B.1.1.529). Proportions of these variants circulating in a human population are typically estimated by large-scale sequencing of individual patient samples. Sequencing a mixture of SARS-CoV-2 RNA molecules from wastewater provides a cost-effective alternative, but requires methods for estimating variant proportions in a mixed sample. From the modeling point of view, a sequenced sample is a pangenome of SARS-CoV-2 strains which needs to be deconstructed into individual genomes. We will briefly explore a solution to this problem and outline limitations of our current approach and some open problems in this area.

Investigating Allelic and Non-Allelic Homologous Recombination through Founder Sequences

Daniel Doerr, University Hospital Düsseldorf, Germany

Homologous recombination is a major driver of genetic variation of popula-

tions. Massive sequencing efforts enable the study of population genetic variation through large collections of genomic sequences, depending on the context called "haplotype reference panel" or "pangenome". In search for compact, descriptive, and computationally amendable representations of pangenomes, the theory of founder sequences has recently celebrated a comeback in the form of (elastic) founder graphs that enable linear time construction and indexability. Yet, founder graphs have limited ability to represent structural variation. The variation graph, an alternative data structure, has gained popularity due to its ability to broadly capture genetic variation, including structural variation. This talk discusses the deep connection between founder and variation graphs with respect to homologous recombination. In particular, we highlight how homologous recombination between non-allelic loci gives rise to structural variation. Consequently, we propose a computational model that unifies both allelic and non-allelic homologous recombination and discuss open problems arising from this model.

List of Participants

- Paola Bonizzoni (Organizer), University of Milano-Bicocca, Italy
- Alberto Policriti (Organizer) University of Udine, Italy
- Kunihiko Sadakane (Organizer) The University of Tokyo, Japan
- Dominik Koeppl, Tokyo Medical and Dental University, Japan
- Nadia Pisanti, University of Pisa, Italy
- Hideo Bannai, Tokyo Medical and Dental University, Japan
- Daniel Doerr, University Hospital Düsseldorf, Germany
- Tomas Vinar, Comenius University, Slovakia
- Simone Ciccolella, University of Milano-Bicocca, Italy
- Fereydoun Hormozdiari, University of California, at David, Italy
- Luca Denti, University of Milano-Bicocca, Italy
- Travis Gagie, Dalhousie University, Canada
- Peter Peresini, Comenius University, Slovakia
- Yuto Nakashima, Kyushu Univeristy, Japan
- Rayan Chikhi, Pasteur Institute, France
- Sung-Hwan Kim, University Ca-Foscari, Italy
- Nicola Prezza, University Ca-Foscari, Italy
- Yoshihiro Shibuya, Institute Pasteur, France
- Sharma Valliyil Thankachan, NC State University, USA

Meeting Schedule

Check-in Day: February 19 (Sun)

- Welcome Banquet

Day1: February 20 (Mon)

- 9.00 - 9.15 Welcoming address – Paola Bonizzoni, Alberto Policriti, Sadakane Kunihiko
- 9.10 - 10.00 Ice-break
- 10.00 - 10.30 Presentation of PANGAIA/ALPACA networks
- 11.15-12.00 Talk session I
- 13.30-15.00 Talk session II

- 15.30-16.30 Talk session III
- Group Photo Shooting

Day2: February 21 (Tue)

- 9.00-11.00 Talk session 4
- 11.15-12.00 Plenary session 1
- 13.30-15.00 Teamwork session 1
- 15.30-16.30 Teamwork session 2

Day3: February 22 (Wed)

- 9.00-11.00 Plenary session 3
- 11.00-12.00 Teamwork session 3
- Excursion and Main Banquet

Day4: February 23 (Thu)

- 9.00-11.00 Plenary session 3
- 11.15-12.00 Teamwork session 4
- 13.30-15.00 Teamwork session 5
- 15.30-16.30 Teamwork session 6

Day5: February 24 (Fri)

- 9.30-11.30 Plenary session 5
- 11.30-12.00 Conclusion and wrap-up

Identified issues and future directions

After the talk sessions, the participants have attended team-group meetings and have been working on specific open problems discussed in the talk sessions. Indeed, the talk sessions have been focused on future research directions. In a plenary session participants have presented specific open problems on which there has been a discussion to finalize a restricted list of topics considered relevant for the scientific community. The following three main topics have been identified as being of common interest:

- haplotype-aware structural variants detection from long sequencing reads in pangenome graphs,
- fast and accurate alignment in textual pangenomes with wavefront-like approaches,
- Mems finding in Wheeler graphs and compressed approximate MEMs finding [1].

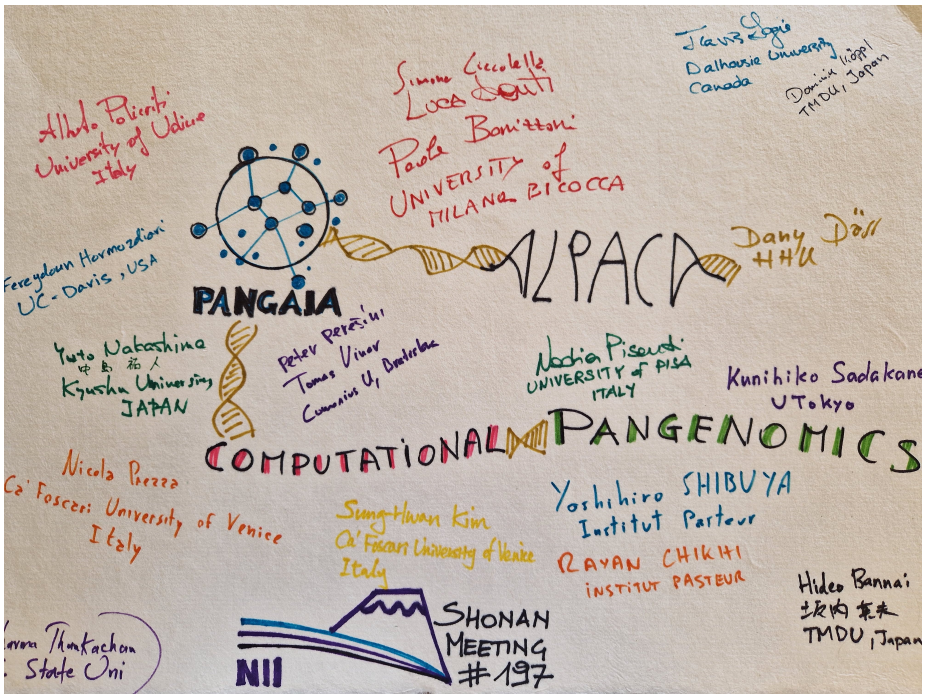

Figure 1: Signs from all participants.

Then the participants have formed three different groups to work on the above topics. Concerning the first topic, some main computational problems have been identified and preliminary results have been drafted related to these computational problems. The team-group of the second topic has discussed about some relevant recent results on wavefront approaches to local and global alignment [2]. A specific research direction for developing an algorithm for local alignment in degenerate texts has been discussed in the meeting. Finally the third group-team has achieved preliminary results on an open problem presented by Nicola Prezza during the talk session. All the advancements reached by the three team-groups have been reported in plenary sessions.The last day of the meeting the organizers have summarized the main contributions and have reported some concluding remarks.

The meeting has pointed out the relevance of notions from combinatorial stringology and automata theory in adopting a new point of view in addressing computational problems such as [3]: building compressed indexes for pangenome graphs (using for example the notion of co-lexicographic ordered automata [4]), identifying structural variations in comparing genomes (as for example the notion of Sample Specific String [5]). On the other hand, compressed data structures are crucial for understanding the complexity of computational problems related to querying pangenome graphs [6], [7]. The meeting has been fruitful in building new collaborations on the above mentioned topics and in identifying new approaches for solving the problem of detecting structural variations in pangenomics.

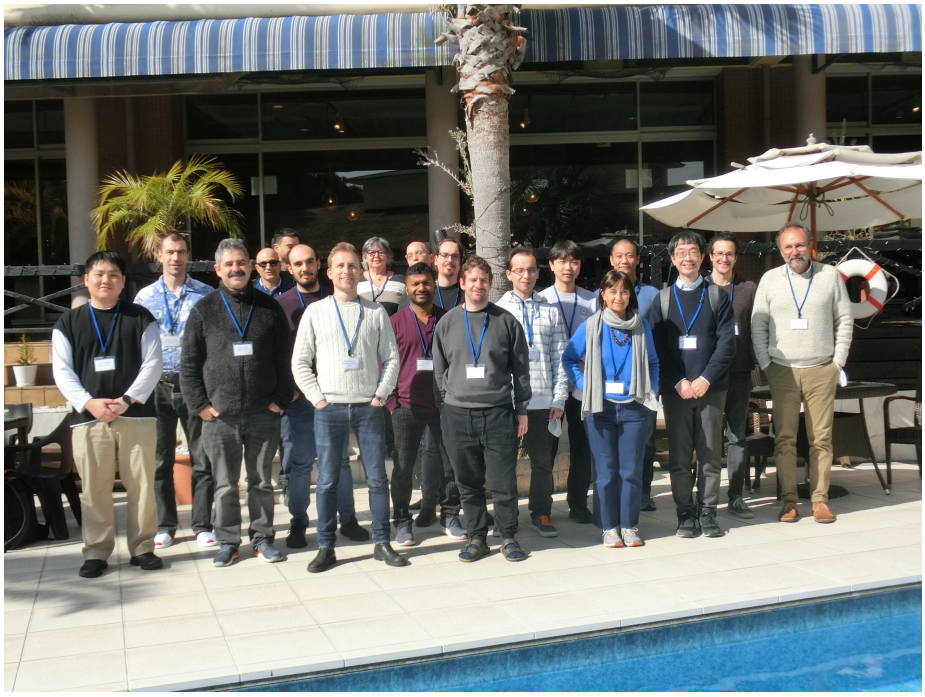

Figure 2: All participants at Shonan OVA

Figure 3: A trip to the beach near Shonan OVA, from where to view Mount Fuji.

Figure 4: The social event: tea ceremony...great experience!

1 Acknowledgements

The organizers thank Shonan for allowing them to organize such a productive meeting. Paola Bonizzoni thanks PANGAIA since she received funding from the European Union's Horizon 2020 Innovative Training Networks programme under the Marie Skłodowska-Curie grant agreement No. 872539. for attending the meeting.

References

[1] Travis Gagie, Giovanni Manzini, and Jouni Sirén. Wheeler graphs: A framework for bwt-based data structures. *Theoretical computer science*, 698:67–78, 2017.

[2] Santiago Marco-Sola, Juan Carlos Moure, Miquel Moreto, and Antonio Espinosa. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*, 37(4):456–463, 2021.

[3] Paola Bonizzoni, Clelia De Felice, Yuri Pirola, Raffaella Rizzi, Rocco Zaccagnino, and Rosalba Zizza. Can formal languages help pangenomics to represent and analyze multiple genomes? In *Developments in Language Theory: 26th International Conference, DLT 2022, Tampa, FL, USA, May 9–13, 2022, Proceedings*, pages 3–12. Springer, 2022.

[4] Giovanna D'Agostino, Nicola Cotumaccio, Alberto Policriti, and Nicola Prezza. On (co-lex) ordering automata. *arXiv preprint arXiv:2106.02309*, 2021.

[5] Luca Denti, Parsoa Khorsand, Paola Bonizzoni, Fereydoun* Hormozdiari*, and Rayan* Chikhi. Svdss: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nature Methods*.

[6] Jasmijn A Baaijens, Paola Bonizzoni, Christina Boucher, Gianluca Della Vedova, Yuri Pirola, Raffaella Rizzi, and Jouni Sirén. Computational graph pangenomics: a tutorial on data structures and their applications. *Natural Computing*, 21(1):81–108, 2022.

[7] Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135, 2018.