Résumés des exposés de

SeqBIM 2019

16 et 17 décembre 2019

Marne-la-Vallée

Graphes de Rauzy et complexité linéaire

Julien Leroy

Résumé

Les graphes de de Bruijn (ou de Rauzy) sont un outil puissant dans l'étude de la combinatoire des mots infinis (ou des systèmes dynamiques symboliques). Ceux-ci fournissent par exemple une suite décroissante de langages réguliers approximant l'ensemble des facteurs d'un mot infini.

Dans cet exposé, nous détaillerons certains concepts qu'ils permettent d'appréhender : mots de retour, sous-groupes du groupe libre, fréquence d'apparition des mots finis, complexité factorielle, représentations adiques,... Nous traiterons en particulier le cas des mots de complexité factorielle linéaire et le cas des mots dendriques.

Recherche de motifs approchés pour l'analyse de longues lectures de séquençage

Hélène Touzet

Résumé

L'algorithmique du texte et l'analyse bioinformatique de séquences biologiques connaissent depuis plusieurs décennies un développement joint mutuellement enrichissant.

Dans cet exposé, nous parlerons en paticulier de la recherche de motifs approchés, et comment ce type de motifs peut aider au traitement de lectures de séquençage bruitées, telles que les longues lectures Nanopore. Les erreurs autorisées par les motifs approchés sont de type insertion, suppression et substitution d'un caractère. Leur description fait appel à des automates, déterministes ou non-déterministes, comme les automates de Levenshtein. Une approche alternative repose sur des découpages sans perte des motifs. Nous décrirons ces différentes méthodes ainsi que les mises en oeuvre efficaces possibles.

Abstract

Survey of sets of k-mer sets data structures for querying large collections of sequencing datasets

Camille Marchet[1], Mikaël Salson[1] and Rayan Chikhi[2]

[1]*BONSAI team, CNRS, Universite de Lille, CRIStAL UMR 9189, Lille, France*
[2]*SeqBio group - Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France*
***Corresponding author**: camille.marchet@univ-lille.fr*

Abstract

The storage, compression and indexing of collections of next-generation sequencing datasets is a fundamental computational problem with many important bioinformatics applications. A natural way to perform those tasks is to use full-text indexes such as the FM-index. However, full-text indexing requires significant computational resources and so far its applications have been limited. In order to enable sequence searches on large databases, data structures improvements were made in pioneer works such as Sequence Bloom Trees [1] or the Bloom Filter Trie [2]. Such works rely on the key idea of storing k-mer set of sets instead of full-text. K-mers sets of sets indexation is linked to the colored De Bruijn Graph intuition that many k-mers can be shared by a majority of screened datasets. One can notice that, since many k-mers are shared across datasets, the representation of sets of sets is not just the concatenation of representation of sets. Redundant information should be factorized and stored efficiently, such as presented in VARI [3].

This field is recently fast-expanding, and includes a range of contributions that goes from theoretical works to more applied research. In this talk we propose a survey of state-of-the-art methods for indexing and querying k-mers set of sets. We present two key features: the storage space of data structures, and the types of queries they support (e.g. search for a given k-mer within many datasets, enumeration of k-mers, dynamic insertions and deletion, etc.). When possible we draw underlying methodological links between them and we propose some common schemes, as an attempt to represent a map of current k-mer set of sets indexation methods.

References

[1] B. Solomon and C. Kingsford. Fast search of thousands of short-read sequencing experiments. *Nature Biotechnology*, 34(3):300–302, 2016.

[2] Guillaume Holley, Roland Wittler, and Jens Stoye. Bloom Filter Trie: an alignment-free and reference-free data structure for pan-genome storage. *Algorithms for Molecular Biology*, 11(1):3, 2016.

[3] Martin D. Muggli, Alexander Bowe, Noelle R. Noyes, Paul S. Morley, Keith E. Belk, Robert Raymond, Travis Gagie, Simon J. Puglisi, and Christina Boucher. Succinct colored de bruijn graphs. *Bioinformatics*, 33(20):3181–3187, 2017.

Abstract

SeqBIM

Impact of the dataset parameters on the quality of long read error correction

Pierre Morisse[1], Thierry Lecroq[2], Arnaud Lefebvre[2]

[1]*Normandie Université, UNIROUEN, INSA Rouen, LITIS, 76000 Rouen, France*
[2]*Normandie Université, UNIROUEN, LITIS, 76000 Rouen, France*
***Corresponding author**: pierre.morisse2@univ-rouen.fr*

Third generation sequencing technologies Pacific Biosciences and Oxford Nanopore Technologies have become widely adopted since their inception in 2011. In particular, their ability to sequence reads reaching tens to thousands of kbps is expected to help solving various problems such has contig and haplotype assembly, scaffolding, and structural variant calling. However, these reads also display high error rates of 10-15% on average, that can even reach up to 30% on old sequencing experiments. Moreover, these errors are mainly composed of insertions and deletions, in contrast to Illumina reads where most errors were substitutions. As a result, long reads require efficient error correction, and a plethora of error correction tools, directly targeted at these reads, were developed in the last few years. These methods can adopt an hybrid approach, using complementary short reads to perform correction, or a self-correction approach, only making use of the information contained in the long reads sequences. Both these approaches make use of various strategies such as multiple sequence alignment ([4, 1]) or de Bruijn graphs ([3, 2]), or even combine different strategies. The dataset parameters, such as coverage, read length, error rate, or sequencing technology, can have an impact on how well a given corrector or a given strategy performs, and can thus drastically reduce the correction quality. We present an in depth benchmark of available long read error correction tools, on a wide variety of datasets, composed of both simulated and real data, with various error rates, coverages, and read lengths, ranging from small bacterial to large mammal genomes.

References

[1] E. Haghshenas, F. Hach, S. C. Sahinalp, and C. Chauve. CoLoRMap: Correcting Long Reads by Mapping short reads. *Bioinformatics*, 32:i545–i551, 2016.

[2] L. Salmela and E. Rivals. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, 30:3506–3514, 2014.

[3] G. Tischler and E. W. Myers. Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv*, 2017.

[4] C. L. Xiao, Y. Chen, S. Q. Xie, K. N. Chen, Y. Wang, Y. Han, F. Luo, and Z. Xie. MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, 14(11):1072–1074, 2017.

Abstract

UMI-VarCal: a UMI-based variant caller for low-frequency variant detection

Vincent Sater[1,2,3]*, Pierre-Julien Viailly[2,3], Thierry Lecroq[1], Philippe Ruminy[2,3], Élise Prieur-Gaston[1], Mathieu Viennot[2,3] and Fabrice Jardin[2,3]

[1]*Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France*

[2]*Centre Henri Becquerel, 76000 Rouen, France*

[2]*Normandie Univ, UNIROUEN, INSERM U1245, Team "Genomics and Biomarkers of Lymphoma and Solid Tumors", 76000 Rouen, France*

*__Corresponding author__: vincent.sater@gmail.com

Abstract

Due to recent advances in the field of oncology, and the use of liquid biopsy to monitor the tumor burden in the blood, the rise of new extremely specific variant calling algorithms has become a must. Sequencing and DNA polymerase errors introduced at low frequencies create new artifactual variants and make the distinction between real variants and artifactual ones a true challenge. However, the recent use of Unique Molecular Identifiers (UMI) in targeted sequencing protocols has offered a trustworthy approach to accurately call low frequency variants. Here, we present UMI-VarCal, a new UMI-based variant caller with remarkably higher specificity compared to other variant callers. Although our variant caller is far from being the only one that uses UMI information to call variants, UMI-VarCal stands out from the crowd by not relying on Samtools to do its pileup. Instead, thanks to an innovative homemade pileup algorithm specifically designed to treat the UMI tags in the reads, our variant caller surpasses the others (SiNVICT [1], DeepSNVMiner [2] and OutLyzer [3]) in terms of specificity. Furthermore, being developed with performance in mind, our tool is considerably more efficient than the other approaches in terms of execution time and memory consumption. We illustrate the results obtained using UMI-VarCal through the sequencing of different samples suffering from lymphoma. We show that UMI-VarCal is faster than raw-reads-based variant callers and more specific than UMI-based ones.

References

[1] C Kockan, F Hach, I Sarrafi, R H Bell, et al. SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. Bioinformatics, Volume 33, Issue 1:26-34, 2017.

[2] Andrews TD, Jeelall Y, Talaulikar D, Goodnow CC, Field MA. DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. PeerJ. 2016;4:e2074, 2016.

[3] Muller E, Goardon N, Brault B, et al. OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice. Oncotarget, 7(48):79485-79493, 2016.

Abstract

Million sequences indexing

Antoine Limasset[1]*

[1]*Univ. Lille, CNRS, Inria, UMR 9189 - CRIStAL.*

***Corresponding author**: antoine.limasset@gmail.com

Abstract

Most methodological contributions handling sequencing data now acknowledge the need to scale up to the terrific throughput that we face nowadays. Since BLAST, a plethora of tools have been developed to handle the massive amount of available reference sequences. Recently, new structures have been proposed to link a short sequence such as a transcript or a gene to sequencing datasets or reference genomes. The challenge of such structures is to be able to index hundreds of thousands of datasets with a reasonable amount of memory while being able to perform fast queries. A prosperous state of the art of efficient tools quickly emerged, SBT SSBT, HowDeSBT, BIGSI, ... based on different combinations of bloom filters in order to link a k-mer to its associated datasets. Those approaches are incredibly efficient: BIGSI was able to index half a million bacterial genomes with 1.5TB. However, not able to scale to all known genomes or all transcriptomes collections. We aim to propose a new data structure that could use an order of magnitude less memory than BIGSI while being able to perform similar queries in terms of accuracy and throughput. Instead of indexing all k-mers of a dataset, we choose to rely on local sensitive hashing methods to index a small subset of the input k-mers. This choice allows the scaling of the methods with a satisfying accuracy on medium-sized queries (1kb or larger). Furthermore, we rely on a matrix-like structure similar to BIGSI, that offer fundamental properties for such an index:

- Constant time insertion of a new reference sequence by adding a new row to the matrix
- Queries rely on reading columns that can be compressed column for lighter structure and faster queries
- Easy to balance structure where memory/accuracy trade-off can be precisely chosen

In this presentation, we show the design of such a structure using the Min-Hash scheme. We present preliminary results on a hundred thousand bacterial genomes on a proof of concept implementation. We compare our performances to BIGSI and Mashscreen, showing that the proposed structure can achieve a comparable accuracy with a better scaling in memory or throughput. We finally discuss the incoming improvements and what can potential pitfalls from this scheme, and its potential applications in mega-scale sequences indexing, clustering, or genome assembly. The open-source implementation is under development and available on Github at https://github.com/Malfoy/Miekki

Abstract

SeqBIM

Recurrence of substitutive Sturmian words

Pablo Rotondo[1]*, Brigitte Vallée[2]

[1] *LITIS, Université de Rouen, France*
[2] *GREYC, CNRS and Université de Caen*

Abstract

Sturmian words are the simplest infinite words that are not periodic, and the recurrence function can be viewed as a waiting time to discover all of the factors of a given length. The probabilistic study of the generic case has been already conducted by the authors in [1] and [2]. The present study focuses on a key subfamily of Sturmian words: the substitutive Sturmian words, built by the application of letter substitutions. Whereas generic Sturmian words are related to general irrational numbers α, substitutive Sturmian words are associated with quadratic irrationals α and forms a discrete subset of the set of all the Sturmian words.

We wish to compare the behaviour of the recurrence function in the two cases –generic Sturmian words and substitutive Sturmian words. In this article we perform the probabilistic analysis of the recurrence function in the case of substitutive Sturmian words, and proves that the distribution of this function behaves in a strongly similar way in both cases – generic Sturmian words and substitutive Sturmian words–.

Even if the results are analog in the two cases, the methods are completely different from the previous article to the present one. Here, we deal with a discrete structure and use tools from Analytic Combinatorics, and generating functions. As shown by Morse and Hedlund in [3], the continued fraction expansion of the slope α plays a central role in the recurrence function, and it is ultimate periodic in the substitutive case. This is why the dynamical system which produces continued fraction expansion plays an important role, via the transfer operators that are associated to this system. They play the role of generating operators that themselves generate the generating functions of interest.

References

[1] Valérie Berthé, Eda Cesaratto, Pablo Rotondo, Brigitte Vallée, and Alfredo Viola. Recurrence function on sturmian words: A probabilistic study. In *Mathematical Foundations of Computer Science 2015*, pages 116–128, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[2] Pablo Rotondo and Brigitte Vallée. The recurrence function of a random Sturmian word. In *2017 Proceedings of the Fourteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 100–114. SIAM, Philadelphia, PA, 2017.

[3] Marston Morse and Gustav A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *Amer. J. Math.*, 62:1–42, 1940.

Abstract

SeqBIM

Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?

Nicolas Guillaudeux[1]*, Catherine Belleannée[1], Samuel Blanquart[1], Jean-Stéphane Varré[2]

[1] *Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France*
[2] *Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France*
***Corresponding author**: nicolas.guillaudeux@inria.fr*

Abstract

In eukaryotic species, the alternative splicing mechanism enables a same gene to express various isoform transcripts. We currently do not know how to determine the whole catalog of isoform transcripts that can be expressed from a gene. Particularly, RNA-seq data allows us to identify only a subset of the expressed transcripts, for technical reasons and due to the low expression levels of some transcripts [1].

To fill in this catalog, we have proposed a comparative genomics method allowing to identify which transcripts known in a source species can also be producted by the orthologous gene of a target genome [2]. This method uses functional sites (start, stop codons and splice sites) known in a given source gene to transpose them, through homology sequence search, into the target orthologous gene. From orthologous exons thus identified, it is then possible to estimate whether a transcript of the source gene has a splicing ortholog, i.e. whether its exon combination can also be expressed by the target gene. The method has been validated on orthologous genes shared by human and mouse [2].

In this work, we adapt the approach to the problem of multi-species comparison by means of graphs of orthologous signals, we apply it to a set of orthologous genes shared in human, mouse and dog, and we predict several thousand of new orthologous transcripts. From that data, we then identify a set of genes conserved in those three species: sharing the same functional sites and expressing the same transcripts. We also identify reasons making some isoform transcripts not feasible. A number of our transcript predictions are confirmed by recent annotation and sequencing data.

References

[1] K. Križanovic, A. Echchiki, J. Roux, and M. Sikic. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics (Oxford, England)*, 34(5):748–754, 2018.

[2] S. Blanquart, J.S. Varré, P. Guertin, A. Perrin, A. Bergeron, and K.M. Swenson. Assisted transcriptome reconstruction and splicing orthology. *BMC Genomics*, 17(786), 2016.

Abstract

SeqBIM

Construction of individual recombination maps using linked-read sequencing data

Andreea Dréau[1]*, Vrinda Venu[2], Elena Avidevich[2], Ludmila Gaspar[2] and Felicity Jones[2]

[1]*MIAT, UR875, INRA, 31326 CASTANET TOLOSAN cedex, France*
[2]*Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany*
***Corresponding author**: andreea.dreau@inra.fr*

Abstract

Meiotic recombination is a major molecular mechanism generating genomic diversity. Recombination rates vary across the genome, often involving localised cross-over "hot-spots" and "cold-spots". Studying the molecular basis of this recombination variation has been challenging due to the expense and data-intensity required to build numerous individualised genome-wide maps of recombination rate.

In this study we introduce a new analysis pipeline called ReMIX [1] that identifies recombination cross-over events across the genome from sperm DNA using linked-read sequencing from 10X Genomics. The linked-read DNA libraries are generated from long DNA molecules trapped inside nanoliter sized droplets. Short reads produced from a given DNA molecule are tagged with a droplet-specific barcode that can be used to computationally reconstruct single molecules after Illumina sequencing. This provides low-cost long-range information facilitating haplotype reconstruction. Applied to DNA from gametes, a small fraction of molecules are spanning meiotic cross-overs. These recombinant molecules are identified by ReMIX as those that switch from one haplotype phase to the other way along the molecule, enabling us to quantify recombination variation across the genome. The linked-read information is exploited by ReMIX during three steps: identification of high-quality heterozygous variants, reconstruction of molecules, and the haplotype phasing of each molecule.

We tested the method by contrasting recombination profiles of gametic and somatic tissue from a hybrid mouse and stickleback fish. Our pipeline faithfully detects previously described recombination hotspots in mice[2] and allows us to characterize and use forward genetic mapping to identify the molecular basis of recombination variation in diverging stickleback species[3].

References

[1] Andreea Dréau, Vrinda Venu, Elena Avdievich, Ludmila Gaspar, and Felicity C Jones. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nature Communications*, 10(4309), 2019.

[2] Kenneth Paigen, Jin P Szatkiewicz, Kathryn Sawyer, Nicole Leahy, Emil D Parvanov, Siemon HS Ng, Joel H Graber, Karl W Broman, and Petko M Petkov. The

recombinational anatomy of a mouse chromosome. *PLoS Genetics*, 4(7):e1000119, 2008.

[3] Marius Roesti, Dario Moser, and Daniel Berner. Recombination in the threespine stickleback genome - patterns and consequences. *Molecular Ecology*, 22(11):3014–3027, 2013.

Extended abstract

SeqBIM

Forbidden substrings and the connectivity of the Hamming graph of RNA sequences: partial disconnectivity tests

Maher Mallem[1], Alain Denise[2]*, Yann Ponty[3]*

[1] *Department of Computer Science, ENS Paris-Saclay, Cachan, France*
[2] *LRI and I2BC, Université Paris-Sud / Paris-Saclay, Gif-sur-Yvette, France*
[3] *LIX, École Polytechnique, Palaiseau, France*
***Corresponding authors**: alain.denise@u-psud.fr and yann.ponty@lix.polytechnique.fr

Abstract

RNA structure design methods have grown in complexity to cover an increasing scope of application. Recent approaches combine an initial random generation with a local optimization step, and consider both a user-specified secondary structure and sets of mandatory and forbidden substrings. Although these additional constraints lead to better design results, they may interfere with the local optimization phase. Indeed, forbidden substrings may disrupt the connectivity of their underlying search space, a key property for the success of the local search. A naive connectivity test would explore the whole graph of candidate sequences, leading to an exponential time connectivity test.

In this work, we propose two partial algorithms based on compact graph structures - the De Bruijn graphs and the Aho-Corasick automaton - allowing the detection of disconnectivity in time independent from the length of RNA sequence. Tested on random instances, our tests were able to detect the disconnectivity with sensitivity ranging between 35% and 55%, motivating further research.

Keywords

RNA Design – Forbidden Substrings – De Bruijn graphs – Aho-Corasick automaton

1. Introduction

First introduced in [1], the computational design of RiboNucleic Acids (RNA) design has been studied extensively over the past decades [2] due to its successful application in a variety of biological contexts [3, 4]. Its ultimate goal is the synthesis of molecules to achieve a targeted biological function. In its simplest form, also called **inverse folding** of RNA, the design problem consists in finding a sequence that adopts a given secondary structure as its Minimum Free Energy (MFE) structure, typically computed using polynomial-time dynamic programming [5]. Given the NP-hardness of the problem [6], recent methods [7, 8, 9, 10] tackle the problem heuristically in two phases: First, an initial **seed sequence** is sampled from a distribution that captures a relaxed version of the objective function [11]; Next, the seed is iteratively refined using a **local search strategy** [1], eventually inducing a Boltzmann-Gibbs distribution with respect to the final objective function (*e.g.* the free-energy difference between the sequence MFE structure and its first suboptimal structure).

However, realistic applications of design require additional **sequence constraints**, for instance to avoid undesired interactions within a cellular context. The seed sampling phase can be adapted to avoid a predefined set $\mathcal{F}$ of **forbidden motifs** using formal language constructs [12] or direct dynamic programming [13]. However, to the best of our knowledge, little to no work has been done to assess the **impact of forbidden motifs on the local search**. Indeed, allowing the local search to violate sequence constraints would lead to very few valid candidate sequences, since an overwhelming proportion of the sequences may (and will, from the monkey/typewriter *paradox*) feature some forbidden motif during the local search.

On the other hand, enforcing the avoidance of $\mathcal{F}$ at each step of the local search may disrupt the **search space connectivity**, or equivalently the non-ergodicity of the Markov Chain induced by the sequence space and the moves set of the local search. For instance, while designing an RNA of length n within an alphabet $\Sigma = \{\mathsf{A}, \mathsf{U}\}$ and $\mathcal{F} = \{\mathsf{AU}, \mathsf{UA}\}$, the only two words avoiding $\mathcal{F}$, A^n and U^n, have Hamming distance n. The search space is thus disconnected for any move set inducing changes of bounded Hamming distance $n' < n$. Such a **disconnectivity** prevents the convergence of the local search, *i.e.* it rules out any (probabilistic) guarantee to ultimately discover promising candidates whenever such candidates exist.

In this work, we address the efficient algorithmic detection of disconnected search spaces for a given set $\mathcal{F}$ of forbidden motifs, a given RNA sequence length n and a given moves set. We restrict our attention to k-Hamming move sets, consisting of symmetric moves $s \leftrightarrow s'$ where both s and s' avoid $\mathcal{F}$, and such that Hamming distance $H(s, s') = k$. A *brute-force* solution would generate the whole search space as a graph, and check the existence of a single connected component in a highly impractical $\mathcal{O}(|\Sigma|^n)$ time complexity. Instead, we exploit the highly-structured nature of the problem to propose partial algorithms, based on the De Bruijn graphs and Aho-Corasick automata, whose complexity depend on $\mathcal{F}$ and k, but remain largely independent from n.

2. Definition of the problem

Let Σ be an alphabet, $|\Sigma| \geq 2$, and $n \in \mathbb{N}, n \geq 2$ be a sequence length. Denote by $\mathcal{F} \subset \Sigma^\star$ the set of forbidden motifs, then $\mathcal{L}_{\mathcal{F},n} \subseteq \Sigma^n$ represents the words that do not contain any motif in $\mathcal{F}$. Let $m(\mathcal{F}) \stackrel{\text{def}}{=} \max_{f \in \mathcal{F}} |f|$, we assume that $n \gg m(\mathcal{F})$.

General problem

Input: Length $n \geq 2$, set $\mathcal{F}$ of forbidden motifs, and neighborhood function $\delta : \mathcal{L}_{\mathcal{F},n} \to \mathcal{L}_{\mathcal{F},n}$
Output: Yes if $G = (\mathcal{L}_{\mathcal{F},n}, \delta)$ is (strongly) connected, No otherwise.

Here we restrict our attention to the k-Hamming neighborhood δ_k for some $k \in [1, n]$, defined for any word $w \in \mathcal{L}_{\mathcal{F},n}$ as $\delta_k(w) = \{w' \in \mathcal{L}_{\mathcal{F}} \mid H(w, w') \leq k\}$ where $H(w, w')$ is the classic Hamming distance between two words w, $w' \in \Sigma^n$.

Since k-Hamming neighborhoods are symmetric, strong connectivity and connectivity are equivalent. The central question, addressed in the following, becomes:

Is the **Hamming graph** $G_{\mathcal{F},n,k} \stackrel{\text{def}}{=} (\mathcal{L}_{\mathcal{F},n}, \delta_k)$ connected?

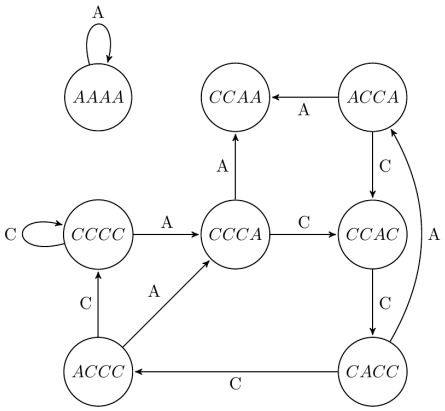


Figure 1. De Bruijn graph $\mathcal{DB}_{\mathcal{F}}$ for $\mathcal{F} = \{\mathsf{ACA}, \mathsf{CAAA}, \mathsf{AAC}\}$ and $\Sigma = \{\mathsf{A}, \mathsf{C}\}$

3. Algorithms

We derive a first partial diconnectivity test from a simple property of De Bruijn graphs. Then using an equivalence relation on the nodes of the De Bruijn graph, we infer a similar partial disconnectivity test on a variant of the Aho-Corasick automaton which is in linear time on the length of the desired sequence.

3.1 Detecting disconnectivity using the De Bruijn graph of $m(\mathcal{F})$-mers

We use variants of the De Bruijn graph [14] to infer the disconnectivity of $G_{\mathcal{F},n,k}$.

Definition 1. *Given a set $\mathcal{F}$ of forbidden motifs, we define:*

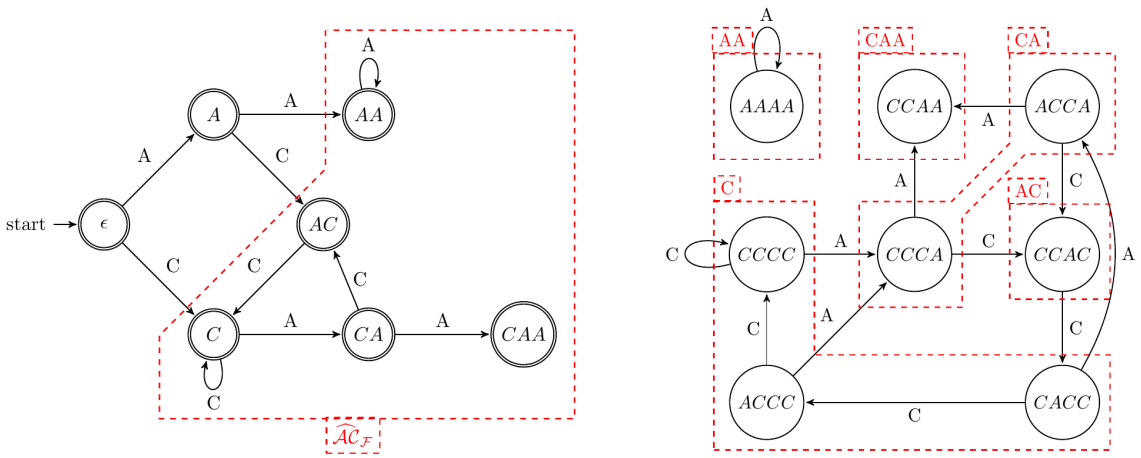
- *The **De Bruijn (di)graph** $\mathcal{DB}_{\mathcal{F}} = (V, E)$ of $\mathcal{F}$, such that $V := \mathcal{L}_{\mathcal{F},m(\mathcal{F})}$, the valid sequences of length $m(\mathcal{F})$, and $E := \left\{(a.w, w.b) \in \mathcal{L}^2_{\mathcal{F},m(\mathcal{F})} \mid a, b \in \Sigma\right\}$;*
- *The **pruned De Bruijn graph** $\mathcal{DB}_{\mathcal{F},n}$, obtained by removing any connected component in $\mathcal{DB}_{\mathcal{F}}$ that cannot generate any word of length n.*

$\mathcal{DB}_{\mathcal{F},n}$ can be built in $\mathcal{O}(|\Sigma|^{m(\mathcal{F})+1})$ time, and detecting unproductive connected components (CC) to build $\mathcal{DB}_{\mathcal{F},n}$ can be done in $\mathcal{O}(|V|)$ time using topological sorting to either detect a cycle ($\rightarrow$ keep CC), or determine n' the length of the longest path ($\rightarrow$ keep CC only if $n' \geq n - m(\mathcal{F}) - 1$).

Remark that $\mathcal{DB}_{\mathcal{F}}$ has $\mathcal{O}(|\Sigma|^{m(\mathcal{F})})$ nodes, and is typically much smaller than the Hamming graph $G_{F,n,k}$ ($\mathcal{O}(|\Sigma|^n)$ nodes), all valid sequences of length n are represented in $\mathcal{DB}_{\mathcal{F}}$ as paths of length $n - m(\mathcal{F})$. For example in Figure 1 the valid sequence CACCAA corresponds to the path $\mathsf{CACC} \rightarrow \mathsf{ACCA} \rightarrow \mathsf{CCAA}$.

Lemma 1. *Upon reading a sequence of letters $a_1.a_2 \dots a_j$, $j \geq m(\mathcal{F})$ from two distinct nodes $u, v \in \mathcal{DB}_{\mathcal{F}}$ the two paths merge at some index $i \leq m(\mathcal{F})$.*

Intuitively, $\mathcal{DB}_{\mathcal{F}}$ can be seen as an automaton, whose states encode the suffixes of length $m(\mathcal{F})$. Thus, after reading $m(\mathcal{F})$ characters the resulting state is $a_1 \dots a_{m(\mathcal{F})}$, irrespectively of the starting state, so the paths either merged at index $m(\mathcal{F})$ or before. This means that if we follow two paths in different connected components of $\mathcal{DB}_{\mathcal{F}}$, the sequence of letters must diverge at least once every $m(\mathcal{F})$ steps, which implies an increasing Hamming distance between the corresponding valid words. This



(a) $\widehat{\mathcal{AC}}_{\mathcal{F}}$ **(b)** $\mathcal{DB}_{\mathcal{F}}$ and the equivalence classes

Figure 2. $\widehat{\mathcal{AC}}_{\mathcal{F},n}$ and $\mathcal{DB}_{\mathcal{F},n}$ when $\mathcal{F} = \{\mathsf{ACA}, \mathsf{AAC}, \mathsf{CAAA}\}$ and $\Sigma = \{\mathsf{A}, \mathsf{C}\}$.

holds for any pair of paths in $\mathcal{DB}_{\mathcal{F}}$ generated from different connected components, leading to the following result.

Theorem 2. *$\forall n \geq (k+1) \times m(\mathcal{F}), \mathcal{DB}_{\mathcal{F},n}$ disconnected $\Rightarrow$ $G_{\mathcal{F},n,k}$ disconnected.*

The implication is not an equivalence, as it is possible to build instances where $G_{\mathcal{F},n,k}$ is disconnected while $\mathcal{DB}_{\mathcal{F},n}$ remains connected. It nevertheless suggests a first algorithm for a partial disconnectivity test within $G_{\mathcal{F},n,k}$: Build $\mathcal{DB}_{\mathcal{F},n}$ and report its connectivity. It has overall time complexity in $\mathcal{O}(|\Sigma|^{m(\mathcal{F})})$, *i.e.* no longer exponential in the sequence length n, yet remains exponential in the length of the forbidden substrings.

3.2 Detecting disconnectivity using the Aho-Corasick automaton of $\mathcal{F}$

Next we attempt to exploit the Nerode equivalence, with respect to the suffix language, of some states in $\mathcal{DB}_{\mathcal{F},n}$.

Definition 3. *Define the Aho-Corasick automaton $\mathcal{AC}_{\mathcal{F}}$ as the DFA having states set $Q = \{u$ proper prefix of some $f \in \mathcal{F}\}$, initial state $q_I = \{\varepsilon\}$, and accepting all words ending in Q. Transitions are $\Delta = \Delta_f \uplus \Delta_b$, with:*

- *Δ_f the forward edges: $\{(u, a, u.a) \mid a \in \Sigma \wedge u, u.a \in Q\}$ (i.e. prefix tree of $\mathcal{F}$)*
- *Δ_b the backward edges: $\{(u, a, v) \mid ua \notin Q \wedge v \in Q$ longest suffix of $u.a\}$*

With this definition of $\mathcal{AC}_{\mathcal{F}}$, a word w is accepted iff no $f \in \mathcal{F}$ is a substring of w, *i.e.* $\mathcal{AC}_{\mathcal{F}}$ recognizes the complement language of the usual Aho-Corasick automaton [15]. Moreover, $\mathcal{AC}_{\mathcal{F}}$ can be built in time $\mathcal{O}(|\Sigma| \times |\mathcal{F}| \times m(\mathcal{F}))$.

Definition 4. *We define:*

- *$\widehat{\mathcal{AC}}_{\mathcal{F}}$ from $\mathcal{AC}_{\mathcal{F}}$ by removing states that are no longer visited after $m(\mathcal{F})$ steps;*
- *$\widehat{\mathcal{AC}}_{\mathcal{F},n}$ as the restriction of $\widehat{\mathcal{AC}}_{\mathcal{F}}$ to components producing words of length n.*

$\lvert\Sigma\rvert$	$m(\mathcal{F})$	n	#Samples	$\#G_{\mathcal{F},n,1}$ discon.	%Rec. $\mathcal{DB}_{\mathcal{F},n}$	%Rec. $\widehat{\mathcal{AC}}_{\mathcal{F},n}$
2	5	10	100 000	36 630	49.5	47.1
2	5	11	100 000	35 893	48.2	46.2
3	5	10	10 000	4 395	53.9	49.2
4	3	6	25 000	9 447	37.6	34.3
4	3	7	10 000	3 728	37.9	35.7
4	4	8	4 000	1 904	54.3	50.1

Figure 3. Recall (TP/P) of our disconnectivity tests for various sets of parameters

As illustrated in Figure 2, grouping together nodes in $\mathcal{DB}_{\mathcal{F}}$ having same prefix/suffix overlaps with forbidden substrings, we get exactly $\widehat{\mathcal{AC}}_{\mathcal{F}}$. This equivalence relation and Theorem 2 imply the following:

Theorem 5. $\forall n \geq (k+1) \times m(\mathcal{F})$, *one has*
$\widehat{\mathcal{AC}}_{\mathcal{F},n}$ *disconnected* $\Rightarrow$ $\mathcal{DB}_{\mathcal{F},n}$ *disconnected* $\Rightarrow$ $G_{\mathcal{F},n,k}$ *disconnected.*

Again, the second implication is only one-way: $\mathcal{DB}_{\mathcal{F},n}$ may be disconnected while $\widehat{\mathcal{AC}}_{\mathcal{F},n}$ remains connected. Still, building $\widehat{\mathcal{AC}}_{\mathcal{F},n}$, and testing its disconnectivity represents an additional partial disconnectivity test for $G_{\mathcal{F},n,k}$. While this variant is expected to detect less cases of disconnectivity, its complexity is significantly better, with the overall construction of $\widehat{\mathcal{AC}}_{\mathcal{F},n}$ now only requiring $\mathcal{O}(|\Sigma| \times |\mathcal{F}| \times m(\mathcal{F}))$ time.

4. Results and Discussion

Both our partial tests were executed on randomly generated sets of forbidden substrings with various parameters. Since the connectivity of the Hamming graph $G_{\mathcal{F},n,k}$ had to be checked on every instance to establish a ground truth, tests could only be conducted with $k = 1$ and small n and $m(\mathcal{F})$ values. The recall (#DetectedDisconnections/#Disconnections, or TP/P) results are given in Figure 3. As expected, the Aho-Corasick-based test always performs slightly worse than the De Bruijn-based one, but not by a large margin ($\sim 5\%$) in our empirical experiments. With a trade-off in accuracy that minimal, the Aho-Corasick-based variant seems to represent a natural first choice in most cases. Recall values range between 35% and 55% for both variants, which is already significant but could probably be improved by exploring subtler relationships between the Aho-Corasick automaton and the Hamming graph.

This preliminary work leaves open several questions of general interest, including:

- What are the shared properties of disconnected instances associated with connected $\widehat{\mathcal{AC}}_{\mathcal{F},n}$? $\mathcal{DB}_{\mathcal{F},n}$?
- Is the problem NP-hard in general?
- How to generalize our constructs to mandatory motifs? To any general automaton generating sequences?
- How to design move sets ensuring connectivity for a given $\mathcal{F}$?

References

[1] Ivo Hofacker, Walter Fontana, Peter Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, Feb 1994.

[2] Alexander Churkin, Matan Drory Retwitzer, Vladimir Reinharz, Yann Ponty, Jérôme Waldispühl, and Danny Barash. Design of RNAs: comparing programs for inverse RNA folding. *Briefings in Bioinformatics*, 19(2):350–358, 01 2017.

[3] Sven Findeiß, Manja Wachsmuth, Mario Mörl, and Peter F Stadler. Design of transcription regulating riboswitches. In *Methods in enzymology*, volume 550, pages 1–22. Elsevier, 2015.

[4] Ryota Yamagami, Mohammad Kayedkhordeh, David H Mathews, and Philip C Bevilacqua. Design of highly active double-pseudoknotted ribozymes: a combined computational and experimental study. *Nucleic acids research*, 47(1):29–42, 2018.

[5] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[6] Édouard Bonnet, Paweł Rzazewski, and Florian Sikora. Designing RNA secondary structures is hard. In *Research in Computational Molecular Biology - 22nd Annual International Conference, RECOMB 2018*, pages 248–250, 2018.

[7] Joseph N Zadeh, Brian R Wolfe, and Niles A Pierce. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry*, 32(3):439–52, 2011.

[8] Matan Drory Retwitzer, Vladimir Reinharz, Yann Ponty, Jérôme Waldispühl, and Danny Barash. incaRNAfbinv: a web server for the fragment-based design of RNA sequences. *Nucleic acids research*, 44(W1):W308–W314, 2016.

[9] Stefan Hammer, Birgit Tschiatschek, Christoph Flamm, Ivo L Hofacker, and Sven Findeiß. RNAblueprint: flexible multiple target nucleic acid sequence design. *Bioinformatics*, 33(18):2850–2858, 04 2017.

[10] Stefan Hammer, Wei Wang, Sebastian Will, and Yann Ponty. Fixed-parameter tractable sampling for RNA design with multiple target structures. *BMC bioinformatics*, 20(1):209, 2019.

[11] Vladimir Reinharz, Yann Ponty, and Jérôme Waldispühl. A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. *Bioinformatics*, 29(13):i308–i315, 2013.

[12] Yu Zhou, Yann Ponty, Stéphane Vialette, Jérôme Waldispühl, Yi Zhang, and Alain Denise. Flexible RNA design under structure and sequence constraints using formal languages. In *ACM-BCB - ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics - 2013*, Bethesda, Washigton DC, United States, September 2013.

[13] Vincent Le Gallic, Alain Denise, and Yann Ponty. Résultats algorithmiques pour le design d'ARN avec contraintes de séquence. In *SeqBio 2015*, pages 26–31, Orsay, France, November 2015.

[14] N. G. De Bruijn. A combinatorial problem. *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, 49:758–764, 1946.

[15] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: An aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June 1975.

SeqBIM

Abstract

Refined upper bounds for the number of designable RNA structures

Hua-Ting Yao[1,2], Cedric Chauve[1,3], Mireille Regnier[1,4], and Yann Ponty[1]

[1] *LIX, École Polytechnique, Palaiseau, France*
[2] *McGill Centre for Bioinformatics, Montréal, Canada*
[3] *Mathematics Department, Simon Fraser University, Burnaby, Canada*
[4] *Inria Lille Nord-Europe, France*
***Corresponding author**: yann.ponty@lix.polytechnique.fr*

The **inverse folding** problem consists in designing an RNA sequence w that adopts a given target structure $S^\star$ as its unique secondary structure of minimum-free energy, with respect to some energy model E. More formally, one must have

$$\underset{S' \text{ comp. with } w}{\operatorname{argmin}} E_{S',w} = \{S^\star\}.$$

Additional design objectives include a min ε value for the **Boltzmann probability**

$$\mathbb{P}(S \mid w) = \frac{e^{-E_{S,w}/RT}}{\mathcal{Z}_w} \geq \varepsilon,$$

with R the Boltzmann constant, T the temperature and $\mathcal{Z}_w := \sum_{S'} e^{-E_{S',w}/RT}$ the partition function; or an upper bound ε' on the **Ensemble defect**, the expected base-pair distance $d(\mathsf{S}, S^\star)$ to $S^\star$ of a random, Boltzmann-distributed, structure S:

$$\mathbb{E}(d(\mathsf{S}, S^\star) \mid w) = \sum_{S' \text{ comp. with } w} \mathbb{P}(S' \mid w) \times d(S^\star, S) \leq \varepsilon'.$$

While apparently diverse, those criteria share a common property: If one cannot be satisfied by any nucleotides assignment for a structure motif M, called a **local obstruction**, then it cannot be satisfied, for any sequence, by any structure that contains M.

Within a recent contribution [1], we have proposed a flexible algorithm to establish a list of 100+ local obstructions within realistic energy models. Counting secondary structures that avoid a set $\mathcal{M}$ of local obstructions is then equivalent to enumerating trees avoiding certain motifs. Using **grammar modeling** and **analytic combinatorics** techniques, we obtain refined asymptotic upper-bounds for the **number of designable secondary structures**, for a variety of design objectives, all of which turn out to be **exponentially smaller** than previously thought.

References

[1] Hua-Ting Yao, Cédric Chauve, Mireille Regnier, and Yann Ponty. Exponentially few RNA structures are designable. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pages 289–298, New York, NY, USA, 2019. ACM.

Abstract

SeqBIM

Efficient single and multiple Cartesian tree matching

Geonmo Gu[1], Siwoo Song[1], Cheol Ryu[1], Simone Faro[2], Thierry Lecroq[3]*, Kunsoo Park[2]

[1] *Seoul National University, Seoul, Korea*
[2] *University of Catania, Catania, Italy*
[3] *Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France*
***Corresponding author**: thierry.lecroq@univ-rouen.fr*

Abstract

Cartesian trees are associated to strings of numbers. They are structured as heap and original strings can be recovered by symmetrical traversal of the trees. Let x be a string of numbers of length m. The Cartesian tree $\mathcal{CT}(x)$ of x is the binary tree where: the root corresponds to the index i of the minimal element of x (if there are several occurrences of the minimal element, the leftmost one is chosen); the left subtree of the root corresponds to the Cartesian tree of $x[1\,..\,i-1]$; the right subtree of the root corresponds to the Cartesian tree of $x[i+1\,..\,m]$.

We are interested in the Cartesian tree matching: given a text and a pattern that consists of sequences of numbers, find all the substrings of the text that have the same Cartesian tree than the pattern. Cartesian tree matching can be applied to finding patterns in time series data such as share prices in stock markets. First solutions for single and multiple Cartesian tree matching appeared in [1]. We present practical solutions for both problems. For the single version we introduce new representations. We present the framework of a binary filtration method and an efficient verification technique for Cartesian tree matching. We also present a SIMD solution for Cartesian tree matching suitable for short patterns. In the case of multiple patterns we present two fingerprinting methods. By combining an efficient fingerprinting method and a conventional multiple string matching algorithm, we can efficiently solve multiple pattern Cartesian tree matching. Experiments show that the proposed algorithms outperform the previous solutions in most cases [2, 3].

References

[1] Sung Gwan Park, Amihood Amir, Gad M. Landau, and Kunsoo Park. Cartesian tree matching and indexing. In *CPM'19*, pages 16:1–16:14, 2019.

[2] Siwoo Song, Cheol Ryu, Simone Faro, Thierry Lecroq, and Kunsoo Park. Fast cartesian tree matching. In *SPIRE'19*, pages 124–137, 2019.

[3] Geonmo Gu, Siwoo Song, Simone Faro, Thierry Lecroq, and Kunsoo Park. Fast multiple pattern cartesian tree matching. *CoRR*, abs/1911.01644, 2019. Accepted to WALCOM'20.

Abstract

A graph-theoretic formulation of the linked-readsbarcode separation problem

Yoann Dufresnde[1]*, Cédric Chauve[2], Rayan Chikhi[1]

[1] *G5 Sequence Bioinformatics, Institut Pasteur, Paris, France*
[2] *Department of Mathematics, Simon Fraser Uiversity, Vancouvert, Canada*
***Corresponding author**: yoann.dufresne@pasteur.fr*

Abstract

Background: Intersection graphs are undirected graphs where each node represents some collection of elements, and edges indicate non-empty intersections between pairs of collections. Interval graphs are special cases of intersection graphs, where nodes represent intervals on the real line. Now consider the following operation: given an interval graph, partition its nodes into independent sets of nodes (i.e. in a partition no two nodes share an edge), and consider the intersection graph of these partitions. We will study the following inverse problem: given an intersection graph produced by the previously described operation, can one recover the original interval graph?

Motivation: This inverse problem turns out to have applications in the analysis of linked-read sequencing data (10X Genomics technology). Long molecules (10-100 kbps) are isolated using microfluidics, and then are cut and sequenced into barcoded short reads. Reads originating from the same molecule have the same barcode. But the converse is not true: a barcode can correspond to more than one molecule. Molecules form a set of genomic intervals, and thus an interval graph. In a reference-free setting, it is not possible to directly construct this interval graph, mainly due to barcode collisions. However, we claim that the intersection graph over barcodes is more readily computable. For the analysis of linked-reads, it would then be desirable to retrieve or approximate the original molecule interval graph (given the barcode intersection graph), and further assign reads to their true molecules. Applications include *de novo* assembly, structural variant detection, and haplotype phasing.

Results: We will present a technique to approximate the molecule (interval) graph given a barcode (intersection) graph. The technique relies on finding d-cliques within the neighborhood of each node in the barcode intersection graph. Then an auxiliary graph is constructed from these d-cliques. It turns out that by traversing the auxiliary graph, one can construct a sequence of barcodes that reflect the consecutive order of their underlying molecules on the genome. We will present preliminary results on simulated data, and discuss ongoing applications on real data.

Abstract

SeqBIM

Longest tandem scattered sub-sequences

Tatiana Rocher[1]* and Luís M. S. Russo[1 2]

[1] *INESC-ID Lisboa, Lisbon, Portugal*
[2] *Department of Computer Science and Engineering, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
***Corresponding author**: tatiana.rocher@tecnico.ulisboa.pt*

Abstract

The problem of the longest tandem scattered sub-sequences aims to find the longest sub-sequence which occurs twice without overlap in a word. Lets consider a partition of a word F: $F = P.S$, where P is a prefix and S is the remaining suffix. To determine which partition yields the overall longest sub-sequence, we need to test all such partitions.

A naive method would compute every longest common subsequences (LCS) of every partition independently. But as there are only two changes between two successive partitions (a letter shifts from the beginning of S to the end of P), we could use previous results.

We use LCS matrices where a matrix uses the computation of the previous one and updates the changes. Every LCS matrix between each partition $P.S$ allow the following updates: the deletion of the last letter of P (last matrix line) and the addition of a letter at the front of S (first column). Only $O(n)$ updates are necessary between two successive partitions. We use a table of Fenwick trees to mimic the LCS matrix and disjoint sets to update the matrix.

This solution solves the problem in $O(n^2\alpha(n^2))$ time, where $\alpha(x)$ is the inverse Ackermann function. Although this solution is no better than the current best known solution [1], the aim is to use this idea for matrices of three and more dimensions in future work.

This work was partly supported by national funds through FCT – Fundação para a Ciência e Tecnologia, under projects NGPHYLO PTDC/CCI-BIO/29676/2017 and UID/CEC/50021/2019.

References

[1] Alexander Tiskin. Semi-local string comparison: Algorithmic techniques and applications. *Mathematics in Computer Science*, 1(4):571–603, 2008.

Abstract

SeqBIM

MinYS: Mine Your Symbiont by targeted genome assembly in symbiotic communities

Cervin Guyomar[1,2,3]*, Wesley Delage[1], Fabrice Legeai[1,2], Christophe Mougel[2], Jean-Christophe Simon[2], Claire Lemaitre[1]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[2]INRA EGI, F-35000 Rennes, France
[3]German Centre for Integrative Biodiversity Research (iDiv), Deutscher Pl. 5E, 04103 Leipzig, Germany
***Corresponding author**: cervin.guyomar@idiv.de

Abstract

The usual approach to recover genomes from metagenomic data consists in a metagenomic assembly of the whole community into contigs, that can then be binned into taxonomic units. This approach is very challenging due to both the many species coexisting in the samples and the polymorphism within these species. However, complete metagenomic assembly seems unnecessary in many biological use-cases, focused on the description of a few key-player organisms, or the analysis of symbiotic organisms within an holobiont. In the present work, we present *MinYS*, an innovative tool for targeted genome assembly from metagenomic reads. *MinYS* relies on two steps, leveraging the benefits of both reference based read recruiting and *de novo* assembly. First, taking advantage of a potentially distant reference genome, a subset of the metagenomic reads is assembled into specific contigs. Then, using an enhanced version of the *MindTheGap* local assembly algorithm, this first draft assembly is completed using the whole metagenomic readset in a *de novo* manner. The resulting assembly can be output as a genome graph, allowing to distinguish different strains with potential structural variants coexisting in the sample. To demonstrate its features, *MinYS* was applied to 50 pea aphid re-sequencing samples in order to recover the genome sequence of its obligatory bacterial symbiont, *Buchnera aphidicola*. *MinYS* was able to return high quality assemblies (one contig complete assembly in more than 90% of samples), even when using distant reference genomes from other aphid hosts. By design, *MinYS* was significantly more time-efficient than full metagenomic assembly to recover this particular genome in the community. In addition, it was able to recover simulated large structural variants, and to return them in a powerful manner thanks to the genome graph representation. As such, it appears as a promising approach for single genome assembly from metagenomic data.

Abstract

SeqBIM

uANI: whole genome comparison and phylogeny reconstruction using sketching

Yoshihiro Shibuya[1]*, Gregory Kucherov[1]

[1]*Laboratoire d'Informatique Gaspard Monge, Université Paris-Est Marne-la-Vallée*
***Corresponding author**: yoshi.itsame@gmail.com

Abstract

Problem: Alignment-free methods for phylogeny reconstruction are usually much faster than whole genome alignments because they do not require to align every base between two sequences to estimate their similarity. On the other hand, with the increasing throughput of sequencing technologies, exact word-based alignment-free methods suffer from high memory usage. For this reason, a broad range of different algorithms have been developed to try to overcome this limitation. Sketching algorithms [1, 2, 3] try to reduce the dimensionality of the input sets while maintaining the ability to compute the similarity measure they are tailored for. Another class of algorithms to speed-up computation are based on sampling the k-mer space to build more manageable algorithms [4, 5, 6]. The minimizer approach is particularly interesting because it allows to sample a Blast reference. The matching minimizers are used as seeds for computing a mapping or a full-fledge alignment. To compute phylogenies, hybrid algorithms use sketches to approximate the similarity of the mapped blocks. In this way, no base-level alignment is needed. A representative tool of this class is FastANI [7] which is specifically tailored to quickly computing ANI (Average Nucleotide Identity) between bacterial genomes.

Results: Inspired by FastANI here we present uANI, an hybrid method that uses count-aware word methods to quickly map and estimate the phylogeny of sequences. uANI proceeds in two steps, first sketching and then the actual comparisons using only the sketches. The sketch is a minimizer index where, for each minimizer, a sketch for the window starting at the same position as the minimizer is also saved. The two are totally independent and it is possible to use different lengths for the minimizers and the sketches. This is particularly important because the window size is usually much shorter than the whole sequence and smaller k-mers are more informative. uANI can be regarded as an ensemble method because it combines multiple recent techniques into one single pipeline. The minimizer index provides rapid pruning of the homology space and approximated Jaccard capability as in [7]. A (suboptimal) chaining algorithm for the windows is used to group together co-linear matches to have meaningful mappings and the counter-based sketches provide much better sensitivity and information for phylogeny reconstruction.

References

[1] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, June 2016.

[2] Yun William Yu and Griffin M. Weber. HyperMinHash: MinHash in LogLog space. *arXiv:1710.08436 [cs]*, July 2019. arXiv: 1710.08436.

[3] Daniel N. Baker and Ben Langmead. Dashing: Fast and Accurate Genomic Distances with HyperLogLog. *bioRxiv*, page 501726, February 2019.

[4] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[5] Hamid Mohamadi, Hamza Khan, and Inanc Birol. ntCard: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics*, 33(9):1324–1330, May 2017.

[6] KDD 2019 | MinJoin: Efficient Edit Similarity Joins via Local Hash Minima.

[7] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):1–8, November 2018.

Abstract

SeqBIM

SVJedi: Structural variations genotyping with long reads

Lolita Lecompte[1]*, Pierre Peterlongo[1], Dominique Lavenier[1] and Claire Lemaitre[1]

[1] *Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France.*
***Corresponding author**: lolita.lecompte@inria.fr*

Abstract

Structural variations (SV) are defined as genomic segments of at least 50 base pair long, that have been rearranged in the genome. Studies on SVs are expanding rapidly. As a result, and thanks to third generation sequencing technologies, more and more SVs are discovered, especially in the human genome. At the same time, for several applications such as clinical diagnoses, it becomes important to genotype newly sequenced individuals on well defined and characterized SVs. Whereas many SV genotypers have been developed for short read data, to our knowledge there is still no approach dedicated to assessing whether some SVs are present or not in a new sequenced sample of long reads, from third generation sequencing technologies.

We present a novel method to genotype known SVs from long read sequencing. The principle of our method is based on the generation of a set of reference sequences that represent the two alleles of each SV. After mapping the long reads to these reference sequences, alignments are analyzed and filtered out to keep only informative ones, to quantify and estimate the presence of each allele. We provide an implementation of this method, SVJedi, available at `https://github.com/llecompte/SVJedi`. Tests on simulated long reads based on 1,000 deletions from the dbVar database show a precision of 97.8 %. On a human genome real Pacific Biosciences dataset for the individual NA24385, 92 % of the genotypes predicted by SVJedi are identical to Genome in a Bottle Consortium's SV benchmark call set.

Abstract

Identification and quantification of strains in a metagenomic sample using variation graphs

Kévin DA SILVA[1,2]*, Nicolas PONS[2], Magali BERLAND[2], Florian PLAZA OÑATE[2], Mathieu ALMEIDA[2], Pierre PETERLONGO[1]

[1]*Univ. Rennes, Inria, CNRS, Irisa*
[2]*MetaGenoPolis, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France*
*__Corresponding author__: kevin.da-silva@inria.fr

Abstract

In the metagenomics field, the classical approach for quantitative analysis of sequencing data consists in aligning sequencing reads against a non-redundant reference gene catalogue that represents a specific ecosystem [1]. However, this approach lacks flexibility and exhaustiveness as it uses a frozen catalogue built from a limited number of samples. To overcome those biases, the reads could be aligned to a more informative reference structure covering the variants encountered in the population. Erik Garrison et al. have developed "*vg*", a toolkit for creating variation graphs, bidirected DNA sequence graphs that represents multiple genomes, including their genetic variation [2]. With a perspective towards metagenomics, we foresee *vg* as a tool enabling to build a catalogue of pangenomes from genomes and metagenomic samples. Our goal is to identify and characterize each strain in terms of gene content and variants by using variation graphs to represent genes families.

We started at genomic level by focusing on *Escherichia coli*, known for its phenotypic diversity (pathogenicity, antibiotics resistance) resulting mostly from its high genomic variability. Genes predicted from genomes of six strains, pathogenic and commensal, were clustered to build a variation graph for each gene family. Among them, the strain O104:H4 was selected as it has been studied during the outbreak of shiga-toxigenic *E. coli* (STEC), which struck Germany in May-June 2011. After mapping reads and through read counting on paths of the graphs, each path corresponding to the successive nodes describing the sequence of a gene, we could identify the mapped strain and its abundance. We will present the results (strains abundance found) on simulated cases and on real data, showing that reads from the German outbreak study can be used to check the STEC-positive and -negative samples using the variation graphs.

References

[1] Li J, et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. 32, 834–841, 2014.

[2] Garrison E, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 36(9):875–9, 2018.