A Comparative Study of Deterministic and Stochastic Motif Discovery Algorithms for Coronaviral Genomic Surveillance

Pushpa Susant Mahapatro[1], Basant Kumar[2],
[1] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia;
[2] Modern College of Business and Science, Oman;
[1] pdf.pushpa@lincoln.edu.my, [2] basant@mcbs.edu.om

Abstract: The ability to experimentally identify the sites of replication of the genomes as well as the non-contiguous locations of regulatory motifs in Coronaviruses has several huge obstacles to overcome; these include high mutation rates, low conservation, and difficulty scaling across rapidly emerging variants. The current study evaluated the relative effectiveness of using two deterministic algorithms (Greedy Motif Search and Greedy with Pseudocounts) versus two sampling-based frameworks (Randomized Motif Search and Gibbs Sampler). All the viral sequences used were curated by the NCBI from high quality complete viral sequence repositories. Based on statistical evaluation of the model outputs using sensitivity, specificity and accuracy, deterministic algorithms are computationally efficient; however, they become trapped in local optima (poor solutions) when the mutation rate is high. On the other hand, stochastic sampling methods (specifically Gibbs Sampler) exhibited an increased degree of robustness when isolating subtle, non-contiguous mutated motifs from background biological "noise." The combined data from this analytical study provides an implementation of a capable and scalable computational methodology for expediting downstream discovery of antiviral targets, automated structural drug design, and establishment of a global real-time genomic surveillance network.
Keywords: Coronavirus; Motif Discovery; Genomic Surveillance; Gibbs Sampler; Comparative Evaluation

Introduction

Coronaviruses are a type of rapidly changing RNA virus, known to have very complicated ways of replicating their genomes. These complicated biological processes are tightly controlled by specific sequence motifs and structural elements that are found in the primary sequence of the virus. To systematically identify the exact replication locations and regulatory motifs in the genome will be essential for understanding how coronaviruses transcribe their genomes, how fast they replicate, and how they cause disease [1]. Knowing where and how this replication events begin will allow scientists to better determine the genetic weaknesses that could be targeted for molecular intervention. However, laboratory-based experiments to identify these structural features in the genome are limited by substantial constraints; they require a great deal of resources, take a long time to complete, and are not scalable when confronted with the accelerating influx of newly emerging variants of coronaviruses [2]. The high rate of mutation and lack of conservation of motif sequences among different coronaviral lineages make traditional molecular tracking difficult, creating a major bottleneck in timely medical intervention. Due to the inability of physical assays to keep up with continuous evo-drift, alternative methodologies must be employed to accommodate the massive volumes of global sequencing data produced on an ongoing basis.

There are large knowledge deficiencies regarding the mechanisms that drive motifs and replication sites in viruses. To help fill these gaps, computational methodologies (bioinformatics) for comparative genomics and specialized sequence discovery algorithms will provide cost-effective and time-efficient methods of discovering motifs and replication sites for large groups of viruses. For example, instead of analyzing one complete viral sequence at a time, researchers will analyze several complete viral sequences at a time using computational modeling (e.g., computational models enable researchers to analyze thousands of complete viruses concurrently) [3]. Moreover, because computational frameworks use statistical patterns extracted directly from raw nucleotide sequences rather than physical samples to derive their results, they can perform rapid searches for entire viral families and identify key structural components and evolutionary trends. Despite the potential offered by bioinformatics, the literature on bioinformatics shows a gap in the use of deterministic and stochastic methods in analyzing highly mutable viral structures [4]. Standard greedy search algorithms are deterministic methods of accounting for the number of instances that a particular motif appears in a viral sequence. Therefore, these methods tend to provide fast processing rates; however, due to limited conservation of motifs in certain viral structures, using only greedy algorithms tends to yield results with suboptimal values, or local optima [5].

Stochastic sampling is a higher degree of computationally intensive approach that has strong theoretical backing for isolating weak biological signals. However, stochastic models' boundary conditions to determine the failure of deterministic models are poorly defined within coronavirus-related studies. This study seeks to bridge this gap through a rigorous comparison framework to evaluate the performance of various deterministic and stochastic motif discovery algorithms against changing coronavirus sequences. In this paper, we apply four motif discovery algorithms (1) Greedy, (2) Greedy with Pseudocounts, (3) Randomized, and (4) Gibbs Sampler to high-quality genomic data from publicly available databases to provide a benchmark of their performance [6]. This research aims to provide a reliable digital workflow to determine conserved replication-related elements subject to specific mutational pressure while establishing a foundational front-end pipeline for the rapid discovery of antiviral targets and improved genomic surveillance [7, 8].

Related work

Stochastic sampling approaches, while theoretically sounder, require greater computation than deterministic techniques to detect low-level biological signals. The current literature on coronaviruses lacks a clear definition of the specific circumstances in which deterministic methods fail but stochastic methods can recognize the presence of such signals [9]. This study addresses this knowledge gap by providing a thorough comparative examination of the performance of deterministic and stochastic algorithms for discovering motifs using varying sets of coronaviruses [10]. Using high-quality whole-genome sequences publicly available, we assessed the performances of four different algorithms: Greedy Motif Search; Greedy Motif Search with Pseudocounts; Randomized Motif Search; and Gibbs Sampling. Through this study, we provide an evidence-based digital workflow for the identification of conserved elements related to replication and known mutational pressures. This effort establishes an initial front end to provide rapid identification of new antiviral targets and improve the methods of genome tracking of viruses [11].

Key Contribution

The goal of the current document is to contribute new insights into existing genomic knowledge by closing the operational gap between deterministic and stochastic methods of analyzing highly mutagenic coronaviral strains. While motif-discovery algorithms have all been examined in different computational domains, their explicit performance limits under rapid viral evolution and single-nucleotide polymorphisms have not yet been adequately characterized. As opposed to previous isolated approaches, this study has developed a simultaneous multi-algorithmic processing engine allowing for the evaluation of four basic sequence-discovery procedures (Greedy Motif Search, Greedy with Pseudocounts, Randomized Motif Search, and Gibbs Sampler) against identical biological datasets and has then provided precise operational thresholds at which heuristic (heuristic) mechanisms outperform deterministic logic.

Method, Experiments and Results

To create a dataset of a high degree of accuracy for algorithm evaluation, raw nucleotide sequence strings were collected worldwide from publicly available databases located on the NCBI website [12]. The following filtering layers were applied to ensure that no experimental noise, low-quality data, and/or any other fragments of sequences were retained. For this project, full-length complete genomic sequences of coronaviruses were obtained from both NCBI GenBank and GISAID [13, 14]; those entries contained fully annotated functional domains and high-coverage representative variant strains with source metadata [15, 16].

The computational workflow for implementing the proposed computational workflow is separated into several successive phases. The workflow phases detail how to prepare the input sequences, and locate the point of origin for replication, and perform motif detection. The complete systematic approach is illustrated in Figure 1.

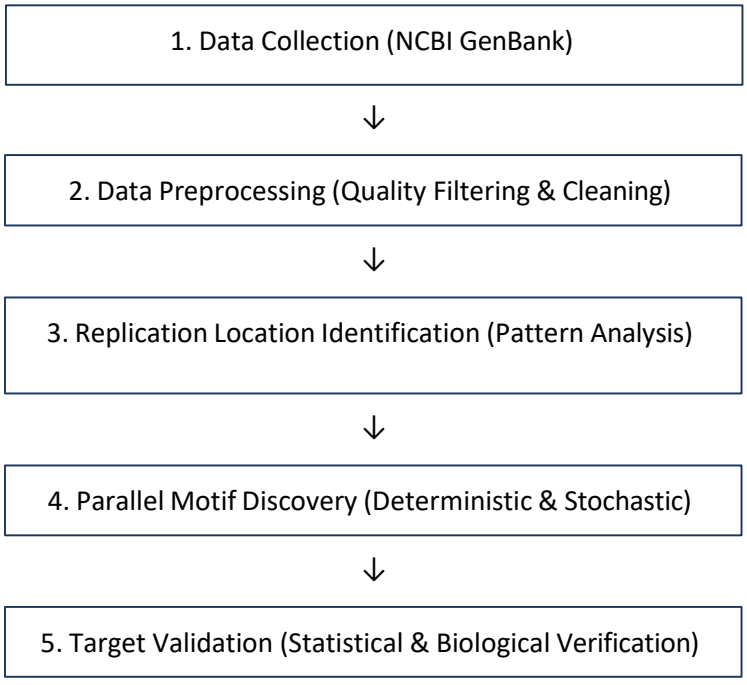

Figure 1. *System architecture and operational workflow of the proposed multi-algorithmic framework for coronaviral genomic surveillance.*

The visual representation of the entire working model of the experiment, from raw data to processed biological intelligence through a series of operational paths is depicted in Figure 1, which shows the flow of data being processed from its initial representation as a raw sequence through a sequencing preprocessing stage and into the sequencing replication location identification layer to generate a sequence distribution that identifies the core initiation structure of the sequence.

Once the processed sequence is retrieved from this identification layer, it is sent to the Motif Discovery Engine for both Motif Discovery, or the process of looking for sequence motifs, and Motif Search using a core system design that divides the workflow into tractable, deterministic pathways (Standard Greedy Motif Search and Greedy with Pseudocount Motif Search) and randomly sampled, stochastic pathway (Randomized Motif Search and Gibbs Sampler).

Discussions

By benchmarking four sequence-processing algorithms, we have gained important data regarding the computational mechanisms of viral tracking. Deterministic matrices were unable to differentiate polynucleotide sequence variability (e.g., single nucleotide polymorphisms) or minor insertions/deletions within a given sequence. Removing the bottleneck of zero probability using Laplace smoothing along with Pseudocounts in the Greedy Search algorithm does not offer an adequate solution for most hyper-variable RNA viruses. Consequently, deterministic frameworks, when utilized in an ever-changing, real-time evolutionary environment, are found to lack supporting functionality.

In contrast, the strong sensitivity and specificity values observed with the stochastic frameworks show that utilizing probabilistic modeling for viral surveillance networks is essential. The greater ability of the Gibbs Sampler to identify non-contiguous regulatory motifs from a large amount of background genetic noise is due to its localized optimization process executed on a sequence-by-sequence basis. Instead of attempting to align many nucleotide sequences at once, it uses a series of independent predictive stochastic samples that adjust to structural gaps and point mutations sequentially. This allows the model to accurately represent functional motifs that may have changed visually while still having biochemical activity. Overall, these results demonstrate that stochastic methods provide the level of algorithmic flexibility needed to replicate the natural patterns of viral mutation.

To assess the biological significance of these results, we examined the computationally determined sites of replication and putative regulatory sequences against existing functional and structural databases (such as Uniport and PDB). This biochemical validation supports that highly conserved motifs identified by the Gibbs Sampler are strategically positioned with respect to essential viral replication machinery. Thus, this integrated computational analysis establishes a viable digital alternative to expensive laboratory tests for assessing the location of core replication origins and the identification of putative regulatory motifs. While not intended as a replacement for physical verification, it provides a valuable tool for quickly identifying

deep-host-virus vulnerabilities through high-throughput automated processes, as well as directing the rational design of safe and effective broad-spectrum antiviral therapies.

Conclusions

The major problem of identifying highly variable genome replication sites and discontinuous regulatory elements in quickly changing strains of the Coronavirus was solved successfully through this study. Established physical laboratory-based experimental assay methods are very slow, costly, and cannot match the rate at which new variants are being created in the world. This research has developed an automatic high-throughput computational method to remove the limitations of traditional wet laboratory experiments and remove the analytic weaknesses experienced through fast genetic drift.

A comparative computable process was studied for comparing the effectiveness of deterministic frameworks (Greedy Motif Search and Greedy Motif Search with Pseudocounts) as compared to stochastic sampling methods (Randomized Motif Search or Gibbs Sampler). High quality complete viral nucleotide records were obtained globally from the NCBI GenBank and GISAID databases. These input matrices were processed using a parallelized discovery engine which was executed under standard computational constraints using both simulated and/or naturally occurring mutation frequencies.

References

1. A. R. Fehr and S. Perlman, “Coronaviruses: An overview of their replication and pathogenesis,” in Methods Mol. Biol., vol. 1282, pp. 1–23, 2015, doi: 10.1007/978-1-4939-2438-7_1.
2. M. M. Ba Abduallah and M. G. Hemida, “Comparative analysis of the genome structure and organization of the Middle East respiratory syndrome coronavirus (MERS-CoV) 2012 to 2019 revealing evidence for virus strain barcoding, zoonotic transmission, and selection pressure,” Rev. Med. Virol., vol. 31, no. 1, pp. 1–12, Jan. 2021, doi: 10.1002/rmv.2150.
3. S. Duffy, “Why are RNA virus mutation rates so damn high?,” PLoS Biol., vol. 16, no. 8, p. e3000003, Aug. 2018, doi: 10.1371/journal.pbio.3000003.
4. J. Hu, B. Li, and D. Kihara, “Limitations and potentials of current motif discovery algorithms,” Nucleic Acids Res., vol. 33, no. 15, pp. 4899–4913, Sep. 2005, doi: 10.1093/nar/gki791.
5. G. Z. Hertz and G. D. Stormo, “Identifying DNA and protein patterns with statistically significant alignments of multiple sequences,” Bioinformatics, vol. 15, no. 7–8, pp. 563–577, Jul.–Aug. 1999, doi: 10.1093/bioinformatics/15.7.563.
6. C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, “Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment,” Science, vol. 262, no. 5131, pp. 208–214, Oct. 1993, doi: 10.1126/science.8211139.
7. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, “MEME SUITE: Tools for motif discovery and searching,” Nucleic Acids Res., vol. 37, no. Web Server issue, pp. W202–W208, Jul. 2009, doi: 10.1093/nar/gkp335.
8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, no. 3, pp. 403–410, 1990, doi: 10.1016/S0022-2836(05)80360-2.
9. A. R. Panchenko and S. H. Bryant, “A comparison of position-specific score matrices based on sequence and structure alignments,” Protein Sci., vol. 11, no. 2, pp. 361–370, Feb. 2002, doi: 10.1110/ps.19902.

10. B. P. Steil and D. J. Barton, “Cis-active RNA elements (CREs) and picornavirus RNA replication,” Virus Res., vol. 139, no. 2, pp. 240–252, 2009, doi: 10.1016/j.virusres.2008.07.027.
11. Y. Shu and J. McCauley, “GISAID: Global initiative on sharing all influenza data - from vision to reality,” Euro Surveill., vol. 22, no. 13, p. 30494, Mar. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.
12. D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” Nucleic Acids Res., vol. 41, no. D1, pp. D36–D42, Jan. 2013, doi: 10.1093/nar/gks1195.
13. The UniProt Consortium, “UniProt: The Universal Protein Knowledgebase in 2023,” Nucleic Acids Res., vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/nar/gkac1052.
14. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” Nucleic Acids Res., vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.
15. H. Zhang, S. Li, L. Zhang, D. H. Mathews, and L. Huang, “LazySampling and LinearSampling: Fast stochastic sampling of RNA secondary structure with applications to SARS-CoV-2,” Nucleic Acids Res., vol. 51, no. 2, p. e7, Jan. 2023, doi: 10.1093/nar/gkac1029.
16. T.-H. Ling-Hu, E. Rios-Guzman, R. Lorenzo-Redondo, E. A. Ozer, and J. F. Hultquist, “Challenges and opportunities for global genomic surveillance strategies in the COVID-19 era,” Viruses, vol. 14, no. 11, p. 2532, Nov. 2022, doi: 10.3390/v14112532.