Communications Theory-Inspired Algorithms for Detecting Protein-Coding Regions in Prokaryotic Genomes: A Comparative Study

Using Communications Theory for Gene Detection in Prokaryotic Genomes: A Comparative Analysis of Correlation-Based Algorithms and Bayesian Classifiers

Mohammad, F., Al Bataineh
Electrical and Communication Engineering Department, UAE University
mffbataineh@uaeu.ac.ae

ABSTRACT

The identification of protein-coding regions in genomic DNA sequences remains a significant challenge in computational genomics, and numerous computational algorithms have been developed to address this problem. Recent advances in this field have facilitated the creation of innovative engineering methods that enable the analysis and modeling of various molecular biology processes. In this work, we propose a novel algorithm for detecting prokaryotic genes that leverages established concepts from communications theory, including correlation, the maximal ratio combining (MRC) algorithm, and filtering techniques, to generate a signal that distinguishes coding and noncoding regions based on characteristic features. The proposed algorithm is applied to multiple prokaryotic genome sequences, and Bayesian classifiers are employed to assess its performance. To further validate the proposed method, we compare its performance to that of established gene detection methods in prokaryotes, such as GLIMMER and GeneMark. This comparison underscores the value of using communications theory concepts for genomic sequence analysis and establishes the efficacy of the proposed algorithm.

CCS CONCEPTS

• **Applied computing** → Life and medical sciences; Genomics; • **Mathematics of computing** → Probability and statistics; Probabilistic inference problems; Bayesian computation.

KEYWORDS

Genomic sequence analysis, Protein-coding regions, Gene detection, Communication theory

ACM Reference Format:
Mohammad, F., Al Bataineh. 2023. Communications Theory-Inspired Algorithms for Detecting Protein-Coding Regions in Prokaryotic Genomes: A Comparative Study: Using Communications Theory for Gene Detection in Prokaryotic Genomes: A Comparative Analysis of Correlation-Based Algorithms and Bayesian Classifiers. In *2023 13th International Conference on Biomedical Engineering and Technology (ICBET 2023), June 15–18, 2023, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3620679.3620687

ICBET 2023, June 15–18, 2023, Tokyo, Japan

ACM ISBN 979-8-4007-0743-8/23/06…$15.00
https://doi.org/10.1145/3620679.3620687

1 INTRODUCTION

The synthesis of proteins in all living organisms is governed by the process of gene expression, which involves two main stages, transcription and translation. Transcription can be understood in the context of coding theory, where the genetic information stored in the deoxyribonucleic acid (DNA) is transformed into the messenger RNA (mRNA). On the other hand, translation can be related to detection theory, where the sequence of mRNA dictates the process of protein synthesis. These two processes are critical for the survival of living organisms as they play a vital role in controlling their structure and development. In this work, we propose a novel algorithm for detecting prokaryotic genes that leverages established concepts from communications theory to overcome limitations of existing methods, such as GLIMMER and GeneMark.

During gene expression, the genetic information encoded in DNA is used to synthesize proteins through two main processes: transcription and translation. Translation is the process by which ribosomes, large complexes of RNA and protein, catalyze the addition of amino acids to a growing polypeptide chain by forming peptide bonds. The ribosome binds to the messenger RNA (mRNA) molecule and scans the nucleotide sequence for a start codon, typically AUG, GUG or UUG, which signals the beginning of the coding region of the mRNA. The ribosome then uses the genetic code to translate the mRNA codons into a sequence of amino acids, which continue to be added to the growing polypeptide chain until a stop codon (UAA, UAG, or UGA) is encountered. At this point, the newly synthesized protein is released from the ribosome and the mRNA. The leader region, composed of nucleotides upstream of the start codon, plays an important role in ribosome binding and initiation of translation.

In genomics, a DNA sequence can be divided into two regions, the genic and intergenic regions, which are also referred to as coding and noncoding regions. A gene is considered the fundamental unit of heredity, consisting of a sequence of nucleotides in the DNA that encodes the synthesis of a gene product, either RNA or protein. The intergenic or noncoding regions refer to DNA sequences that are located between genes. Interpretation of the nature and role of genomic sequence structure present in both the coding and

noncoding regions is a substantial target in genomic research. Gene identification is a crucial step in accomplishing this target, allowing for a better understanding of the functional significance of the DNA sequence.

Gene prediction, also known as gene finding, is a crucial task in identifying regions of genomic DNA that encode genes, including protein-coding and RNA genes, as well as other functional elements such as regulatory regions. In the literature, a plethora of methods have been proposed for gene identification in prokaryotes, such as probabilistic methods [1], [2], statistical methods [3–5], machine learning [6], [7], free energy calculations [8], support vector machines [9], Bayesian methods [10], information theory [11], signal processing methods [11][12], hidden Markov models such as GeneMark [13–19], and interpolated Markov models such as GLIMMER [20]. These methods utilize various techniques, including statistical modeling, machine learning, and signal processing, to detect the presence of genes in genomic sequences. Despite their differences, the ultimate goal of these methods is to accurately identify the location and structure of genes, which is essential for understanding the biological function of the genome.

This work presents a new approach for detecting genes in prokaryotic genomes based on fundamental concepts and principles in communications theory and digital signal processing. The proposed method utilizes several mapping systems to obtain numerical representations of the genomic sequences being studied. By exploiting the concepts of correlation, filtering, and maximum ratio combining, this algorithm can accurately identify coding and noncoding regions in prokaryotic genomic sequences, such as those found in E. coli bacteria. More specifically, this algorithm relies on the identification of a particular property that is commonly found in protein-coding regions, known as the period-3 structure. This periodicity reveals correlations between nucleotide positions along coding sequences caused by the asymmetry in genomic structure at the three coding positions. In contrast to our previous work [21], this paper employs an expanded set of eleven numerical representations of the genomic sequence understudy and extends the correlation task to include sixty-four hypothetical sequences that exhibit the period-3 structure. This enhanced and more generalized version of our previous work improves the methodology, performance, and experimental validation of our proposed algorithm.

To further validate the obtained simulation results, a Bayesian classification technique is applied to provide more precise performance measures, including sensitivity and specificity. Several classification variables were investigated to optimize the algorithm's performance. The results obtained through simulations outperformed those of well-known prokaryotic gene detection algorithms, such as GLIMMER and GeneMark, in terms of sensitivity and specificity. These findings further support the significance of utilizing concepts from communications theory for genomic signal processing.

The remainder of this paper is organized as follows: Section 2.1 details the numerical representations utilized for the genomic sequences under investigation. Section 2.2 presents the mathematical formulation of the proposed gene identification algorithm. In Section 3, we present the simulation results and performance evaluation of the algorithm, including a comparison to existing methods, such as GLIMMER and GeneMark. Lastly, in Section 4, we provide concluding remarks and recommendations for future research.

2 THE PROPOSED GENE IDENTIFICATION ALGORITHM

This Section lists all the numerical representations for the genomic sequences understudy and provides a mathematical description of the proposed gene identification algorithm.

2.1 Numerical Representations

In order to apply digital signal processing techniques to genomic sequences, an appropriate numerical representation is needed. This study proposes several numerical representations for the input genomic sequence, from which the best representation is selected for gene detection via Bayesian classification. Table 1 provides a description of the 12 different numerical representations used in this work. The fourth column of Table 1 applies these representations to the genomic sequence g[n] composed of the nucleotide bases [A,T,C,G], producing a resulting sequence x[n]. By applying these numerical representations, the proposed algorithm is able to detect and distinguish coding and noncoding regions in prokaryotic genomic sequences.

2.2 Mathematical Formulation

The proposed prokaryotic gene detection algorithm consists of several steps. Firstly, the entire prokaryotic DNA sequence is scanned to identify all possible open reading frames (ORFs) that meet specific criteria, including length and initiation and termination codons. The choice of these criteria is based on the biological properties of prokaryotic genes, which typically exhibit distinct features in their coding regions. The identified ORFs are then mapped using one of twelve possible numerical representations proposed in Table 1. These numerical representations were selected to capture different aspects of the genomic sequence, enhancing the ability of our algorithm to detect and distinguish coding and noncoding regions. Hypothetical genomic sequences that exhibit a clear period-3 structure are then generated, and these are mapped to the numerical representation selected in the previous step. The ORFs are then correlated with all possible hypothetical period-3 sequences (64 sequences), and the resulting output sequences are combined using a maximum ratio combining module. The use of maximum ratio combining allows for better signal-to-noise ratio in the output, improving the detection of coding regions. The combined output is then passed through a period-3 filter, and a predefined metric is assigned to every possible input ORF based on the filtered output sequence. The period-3 filter is used to enhance the periodic features of coding regions, making them more distinguishable from noncoding regions. The resulting metric is used as a classification variable for the next stage of Bayesian classification. The Bayesian classifiers are employed to assess the performance of the algorithm by using the metric to calculate the posterior probability of an ORF being a coding region. This allows for a more robust classification of coding and noncoding regions, improving the overall accuracy of our method. The algorithm output metric will be used to determine the validity of the input ORF as a protein-coding sequence. Figure 1 shows a diagrammatical representation of the proposed algorithm.

Table 1: Proposed numerical representations.

No.	Name	Numeric Representation	
1	Integer	$x[n] = \begin{cases} 0 & , \ g[n] = T \\ 1 & , \ g[n] = C \\ 2 & , \ g[n] = A \\ 3 & , \ g[n] = G \end{cases}$.	$[2, 0, 1, 3]$
2	Real	$x[n] = \begin{cases} 1.5 & , \ g[n] = T \\ 0.5 & , \ g[n] = C \\ -1.5 & , \ g[n] = A \\ -0.5 & , \ g[n] = G \end{cases}$	$[-1.5,\ 1.5,\ 0.5,\ -0.5]$
3	Complex	$x[n] = \begin{cases} 1-j & , \ g[n] = T \\ -1+j & , \ g[n] = C \\ 1+j & , \ g[n] = A \\ -1-j & , \ g[n] = G \end{cases}$	$[1+\mathrm{j},\ 1-j, -1+\mathrm{j},\ -1-j]$
4	Polyphase	$x[n] = \begin{cases} -j & , \ g[n] = T \\ j & , \ g[n] = C \\ 1 & , \ g[n] = A \\ -1 & , \ g[n] = G \end{cases}$	$[1+\mathrm{j},\ 1-j, -1+\mathrm{j},\ -1-j]$
5	Quaternion	$x[n] = \begin{cases} -i+j-k & , \ g[n] = T \\ i-j-k & , \ g[n] = C \\ i+j+k & , \ g[n] = A \\ -i-j+k & , \ g[n] = G \end{cases}$	$[i+j+k, -i-j-k,$ $i-j-k,\ -i-j+k]$
6	EIIP	$x[n] = \begin{cases} 0.1335 & , \ g[n] = T \\ 0.1340 & , \ g[n] = C \\ 0.1260 & , \ g[n] = A \\ 0.0806 & , \ g[n] = G \end{cases}$	$[0.1260,\ 0.1335,$ $0.1340,\ 0.0806]$
7	Atomic Number	$x[n] = \begin{cases} 66 & , \ g[n] = T \\ 58 & , \ g[n] = C \\ 70 & , \ g[n] = A \\ 78 & , \ g[n] = G \end{cases}$	$[70, 66, 58, 78]$
8	Paired Numeric	$x[n] = \begin{cases} 1 & , \ g[n] = T \\ -1 & , \ g[n] = C \\ 1 & , \ g[n] = A \\ -1 & , \ g[n] = G \end{cases}$	$[1,\ 1,\ -1,\ -1]$
9	Voss	$x_i[n] = \begin{cases} 1 & , \ g[n] = i \\ 0 & , \ Otherwsie \end{cases}$ $i \in \{A,\ T,\ C,\ G\}$	$x_A[n] = [1,\ 0,\ 0,\ 0]$ $x_T[n] = [0,\ 1,\ 0,\ 0]$ $x_C[n] = [0,\ 0,\ 1,\ 0]$ $x_G[n] = [0,\ 0,\ 0,\ 1]$
10	Tetrahedron	$x_r[n] = \begin{cases} \frac{2\sqrt{2}}{3} & , \ g[n] = T \\ \frac{-\sqrt{2}}{3} & , \ g[n] = \{C,\ G\} \\ 0 & , \ Otherwise \end{cases}$ $x_r[n] = \begin{cases} \frac{\sqrt{6}}{3} & , \ g[n] = C \\ \frac{-\sqrt{6}}{3} & , \ g[n] = G \\ 0 & , \ Otherwise \end{cases}$ $x_b[n] = \begin{cases} 1 & , \ g[n] = A \\ -\frac{1}{3} & , \ Otherwsie \end{cases}$	$x_r[n] = [0,\ \frac{2\sqrt{2}}{3},\ \frac{-\sqrt{2}}{3},\ \frac{-\sqrt{2}}{3}]$ $x_r[n] = [0,\ 0,\ \frac{\sqrt{6}}{3},\ \frac{-\sqrt{6}}{3}]$ $x_r[n] = [0,\ \frac{-1}{3},\ \frac{-1}{3},\ \frac{-1}{3}]$
11	DNA Walk	$x[n] = \begin{cases} x[n-1]+1 & , \ g[n] = \{C, T\} \\ x[n-1]-1 & , \ g[n] = \{A, G\} \end{cases}$	[-1, 0, 1, -1]
12	Z-Curve	$x_1[n] = \begin{cases} x[n-1]+1 & , \ g[n] = \{T, G\} \\ x[n-1]-1 & , \ g[n] = \{A, C\} \end{cases}$ $x_2[n] = \begin{cases} x[n-1]+1 & , \ g[n] = \{A, C\} \\ x[n-1]-1 & , \ g[n] = \{T,\ G\} \end{cases}$ $x_3[n] = \begin{cases} x[n-1]+1 & , \ g[n] = \{A,\ T\} \\ x[n-1]-1 & , \ g[n] = \{C,\ G\} \end{cases}$	$x_1[n] = [-1,\ 0,\ -1,\ 0]$ $x_2[n] = [1,\ 0,\ 1,\ 0]$ $x_3[n] = [1,\ 2,\ 1,\ 0]$

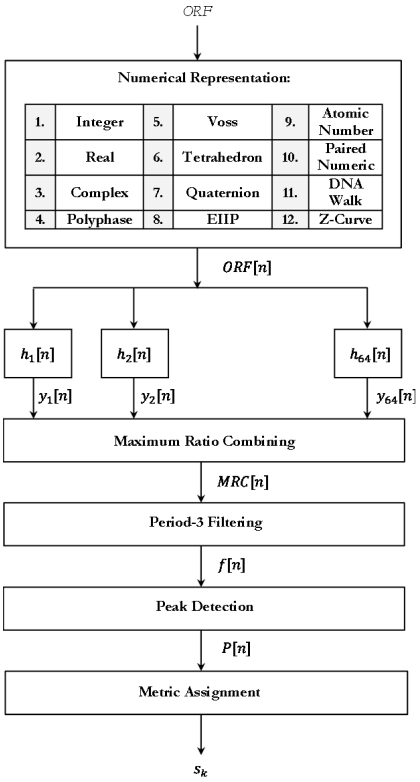

Figure 1: The proposed Gene Detection Algorithm.

3 SIMULATION RESULTS

The proposed prokaryotic gene detection algorithm was evaluated on the complete DNA sequences of five Escherichia Coli Bacterial strains, as listed in Table 2. The algorithm's performance was compared to two well-known prokaryotic gene-finding software, GLIMMER and GeneMark, as shown in Table 2. The comparison includes the number of true positives (TPs), false positives (FPs), false negatives (FNs), and true negatives (TNs) predicted by each method for each of the five genomes. Additionally, the table presents the sensitivity, specificity, correlation coefficient (CC), and approximate correlation coefficient (AC) of the proposed algorithm compared to GLIMMER and GeneMark for each of the five genomes.

A key observation from Table 2 is that the proposed Period-3 Classifier consistently demonstrated higher sensitivity values compared to GLIMMER and GeneMark across all five genomes. This suggests that the proposed algorithm is more effective in detecting true protein-coding regions. However, the specificity values show some variability across the different genomes, indicating that the algorithm's performance in distinguishing non-coding regions could be further improved.

The simulation results demonstrate the effectiveness of the proposed algorithm, with performance metrics that consistently outperformed or were competitive with GLIMMER and GeneMark. The simulation results presented in Table 2 provide strong evidence for the significance and biological relevance of the proposed prokaryotic gene detection algorithm. Notably, these results were obtained using the complex mapping method (No. 3 in Table 1) for numerical representation of the genomic sequences under investigation, further verifying the effectiveness of this approach. While further testing is needed to assess the other suggested mappings, it is worth mentioning that the EIIP mapping performed comparably well to the complex mapping.

However, it is important to acknowledge the limitations of the current study. The algorithm was tested on a limited number of bacterial genomes, and the performance may vary when applied to a broader range of prokaryotic organisms. Furthermore, the algorithm's reliance on period-3 structure identification might not be suitable for all types of prokaryotic genes.

4 CONCLUSIONS

In this study, well-established concepts in communications theory, including correlation, the maximal ratio combining (MRC) algorithm, and filtering techniques, were utilized for prokaryotic gene

Table 2: Performance Evaluation

Genome	Method	Metrics (TP, FP, FN, TN)	Sens.	Spec.	(AC)	(CC)
MG1655	GLIMMER	3561, 915, 580, 434618	0.8599	0.9979	0.826	0.8254
MG1655	GeneMark	3683, 694, 458, 434839	0.8894	0.9984	0.8641	0.8638
MG1655	Period-3 Classifier	3858, 1000, 342, 434816	0.9186	0.9977	0.8548	0.8526
O157:H7	GLIMMER	4133, 1450, 1092, 512544	0.791	0.9972	0.7632	0.7628
O157:H7	GeneMark	4246, 1135, 979, 512859	0.8126	0.9978	0.7988	0.7987
O157:H7	Period-3 Classifier	4350, 1225, 930, 311865	0.8239	0.9961	0.7986	0.7983
Salmonella Typhimurium LT2	GLIMMER	3569, 860, 587, 412831	0.8588	0.9979	0.8305	0.8301
Salmonella Typhimurium LT2	GeneMark	3712, 642, 444, 413049	0.8932	0.9984	0.8715	0.8713
Salmonella Typhimurium LT2	Period-3 Classifier	3819, 751, 681, 39533	0.8487	0.9814	0.8244	0.8244
Staphylococcus Aureus Mu50	GLIMMER	2345, 358, 292, 314445	0.8893	0.9989	0.8774	0.8773
Staphylococcus Aureus Mu50	GeneMark	2407, 316, 230, 314487	0.9128	0.999	0.8975	0.8974
Staphylococcus Aureus Mu50	Period-3 Classifier	2630, 960, 320, 313850	0.8915	0.997	0.81	0.8062
Bacillus Subtilis	GLIMMER	3569, 860, 587, 412831	0.8588	0.9979	0.8305	0.8301
Bacillus Subtilis	GeneMark	3712, 642, 444, 413049	0.8932	0.9984	0.8715	0.8713
Bacillus Subtilis	Period-3 Classifier	3949, 656, 251, 413242	0.9402	0.9984	0.8978	0.8969

detection. The proposed algorithm is based on the period-3 structure present in protein-coding regions of prokaryotic genomes. The algorithm was applied to the genomic sequences of five bacterial strains, including MG1655, O157:H7, Salmonella Typhimurium LT2, Staphylococcus Aureus Mu50, and Bacillus Subtilis. Bayesian classifiers were designed to evaluate the algorithm's performance, and the simulation results demonstrated superior sensitivity and specificity when compared to well-known gene detection methods in prokaryotes such as GLIMMER and GeneMark. Notably, the Period-3 Classifier showed consistently higher sensitivity values, indicating its effectiveness in detecting true protein-coding regions. These findings support the relevance and significance of utilizing concepts from communications theory in genomic sequence analysis.

While the proposed algorithm shows promise, it is important to acknowledge its limitations. The algorithm was tested on a limited number of bacterial genomes, and its performance may vary when applied to a broader range of prokaryotic organisms. Furthermore, the algorithm's reliance on period-3 structure identification might not be suitable for all types of prokaryotic genes.

Future work should focus on further optimization and exploration of additional numerical representations for improved performance. Moreover, the algorithm could be adapted for use in other genomics research areas, such as transcriptome or epigenome analysis, and even extended to eukaryotic organisms.

REFERENCES

[1] T. Yada, Y. Totoki, T. Takagi, and K. Nakai, "A novel bacterial gene-finding system with improved accuracy in locating start codons," DNA Res., vol. 8, no. 3, pp. 97–106, 2001.

[2] J. Besemer, A. Lomsadze, and M. Borodovsky, "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions," Nucleic Acids Res., vol. 29, no. 12, pp. 2607–2618, 2001.

[3] M. Walker, V. Pavlovic, and S. Kasif, "A comparative genomic method for computational identification of prokaryotic translation initiation sites," Nucleic Acids Res., vol. 30, no. 14, pp. 3181–3191, 2002.

[4] S. S. Hannenhalli, W. S. Hayes, A. G. Hatzigeorgiou, and J. W. Fickett, "Bacterial start site prediction," Nucleic Acids Res., vol. 27, no. 17, pp. 3577–3582, 1999.

[5] T. Nishi, T. Ikemura, and S. Kanaya, "GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences," Gene, vol. 346, pp. 115–125, 2005.

[6] W. S. Hayes and M. Borodovsky, "How to interpret an anonymous bacterial genome: machine learning approach to gene identification," Genome Res., vol. 8, no. 11, pp. 1154–1171, 1998.

[7] J. Yu *et al.*, "Prediction of protein-coding small ORFs in multi-species using integrated sequence-derived features and the random forest model," Methods, vol. 210, no. January, pp. 10–19, 2023, doi: 10.1016/j.ymeth.2022.12.003.

[8] Y. Osada, R. Saito, and M. Tomita, "Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes," Bioinformatics, vol. 15, no. 7, pp. 578–581, 1999.

[9] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller, "Engineering support vector machine kernels that recognize translation initiation sites," Bioinformatics, vol. 16, no. 9, pp. 799–807, 2000.

[10] E. M. Crowley, "A Bayesian method for finding regulatory segments in DNA," Biopolymers, vol. 58, no. 2, pp. 165–174, 2001.

[11] T. D. Schneider, "Measuring molecular information," J. Theor. Biol., vol. 201, no. 1, pp. 87–92, 1999.

[12] M. Al Bataineh and Z. Al-qudah, "A novel gene identification algorithm with Bayesian classification," Biomed. Signal Process. Control, vol. 31, 2017, doi: 10.1016/j.bspc.2016.07.002.

[13] J. Besemer and M. Borodovsky, "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses," Nucleic Acids Res., vol. 33, no. suppl 2, pp. W451–W454, 2005.

[14] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a Hidden Markov Model," J. Comput. Biol., vol. 4, no. 2, pp. 127–141, 1997.

[15] R. Raman and G. C. Overton, "Application of hidden markov modeling in the characterization of transcription factor binding sites," Proc. Twenty-Seventh Annu. Hawaii Int. Conf. Syst. Sci., vol. 5, pp. 275–283, 1994.

[16] A. Krogh, I. S. Mian, and D. Haussler, "A Hidden Markov Model That Finds Genes in Escherichia-Coli DNA," Nucleic Acids Res., vol. 22, no. 22, pp. 4768–4778, 1994.

[17] S. R. Eddy, "Hidden Markov models," Curr. Opin. Struct. Biol., vol. 6, no. 3, pp. 361–365, 1996.

[18] S. R. Eddy, "Hidden Markov models and genome sequence analysis," Faseb J., vol. 12, no. 8, pp. A1327–A1327, 1998.

[19] P. Techa-Angkoon, K. L. Childs, and Y. Sun, "GPRED-GC: A Gene PREDiction model accounting for 5 '- 3' GC gradient," BMC Bioinformatics, vol. 20, Dec. 2019, doi: 10.1186/s12859-019-3047-3.

[20] A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg, "Identifying bacterial genes and endosymbiont DNA with Glimmer," Bioinformatics, vol. 23, no. 6, pp. 673–679, 2007.

[21] M. Al Bataineh, "Identification of Coding Regions in Prokaryotic DNA Sequences Using Bayesian Classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12108 LNBI, pp. 3–14, 2020, doi: 10.1007/978-3-030-45385-5_1.