

Volume 8 ▪ Issue 1 ▪ August 2014

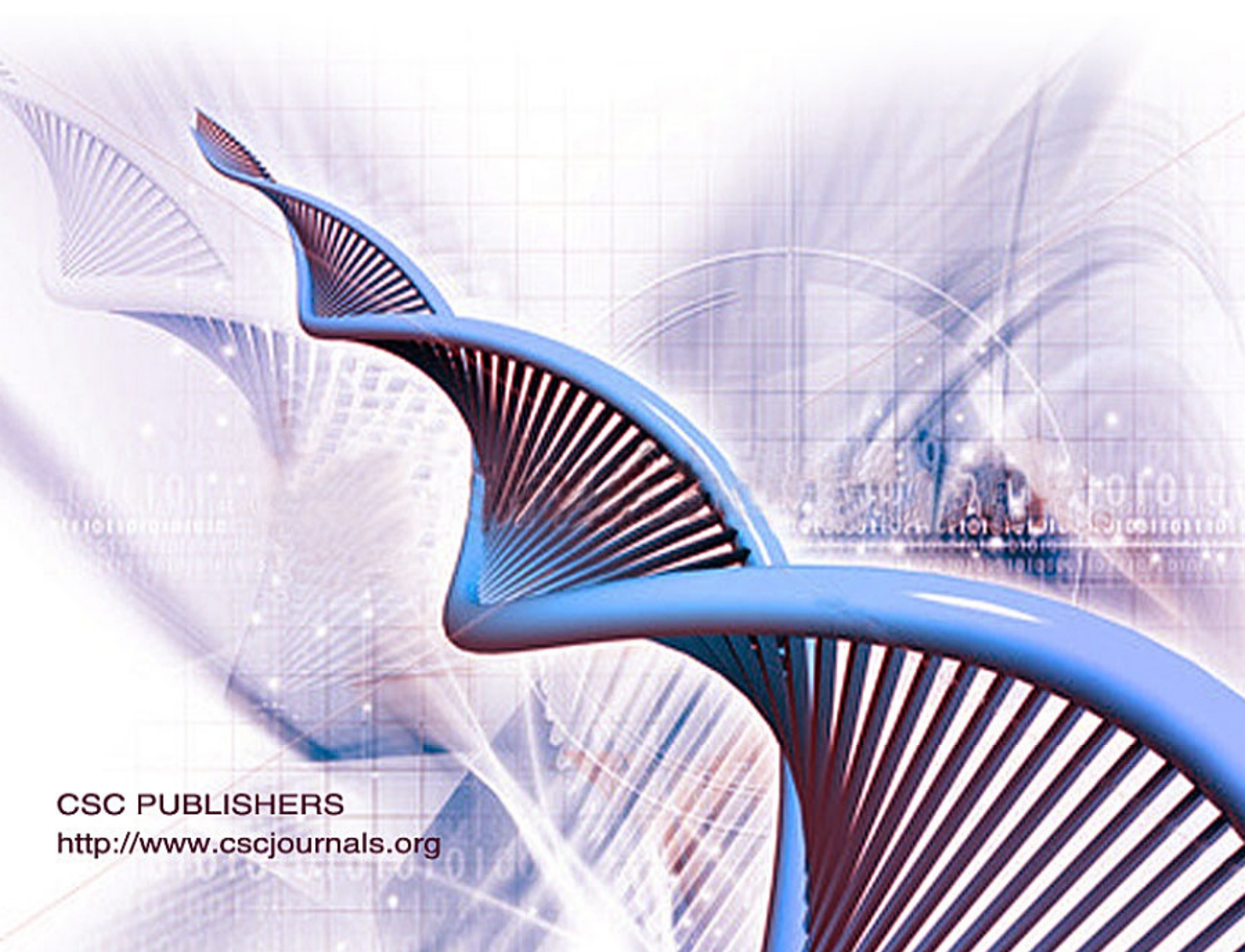
Editor-in-Chief  
Professor João Manuel R. S. Tavares

INTERNATIONAL JOURNAL OF

# BIOMETRICS AND BIOINFORMATICS (IJBB)

ISSN : 1985-2347

Publication Frequency: 6 Issues / Year



CSC PUBLISHERS  
<http://www.cscjournals.org>

# **INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)**

**VOLUME 8, ISSUE 1, 2014**

**EDITED BY  
DR. NABEEL TAHIR**

ISSN (Online): 1985-2347

International Journal of Biometrics and Bioinformatics (IJBB) is published both in traditional paper form and in Internet. This journal is published at the website <http://www.cscjournals.org>, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJBB Journal is a part of CSC Publishers

Computer Science Journals

<http://www.cscjournals.org>

# **INTERNATIONAL JOURNAL OF BIOMETRICS AND BIOINFORMATICS (IJBB)**

Book: Volume 8, Issue 1, August 2014

Publishing Date: 10-08-2014

ISSN (Online): 1985-2347

This work is subjected to copyright. All rights are reserved whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication of parts thereof is permitted only under the provision of the copyright law 1965, in its current version, and permission of use must always be obtained from CSC Publishers.

IJBB Journal is a part of CSC Publishers

<http://www.cscjournals.org>

© IJBB Journal

Published in Malaysia

Typesetting: Camera-ready by author, data conversion by CSC Publishing Services – CSC Journals, Malaysia

**CSC Publishers, 2014**

## EDITORIAL PREFACE

This is the *First* Issue of Volume *Eight* of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics is not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with Volume 8, 2014, IJBB appears with more focused issues related to biometrics and bioinformatics studies. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

### **Editorial Board Members**

International Journal of Biometric and Bioinformatics (IJBB)

## **EDITORIAL BOARD**

### **EDITOR-in-CHIEF (EiC)**

**Professor João Manuel R. S. Tavares**  
University of Porto (Portugal)

### **ASSOCIATE EDITORS (AEiCs)**

---

**Assistant Professor. Yongjie Jessica Zhang**  
Mellon University  
United States of America

**Professor. Jimmy Thomas Efirid**  
University of North Carolina  
United States of America

**Professor. H. Fai Poon**  
Sigma-Aldrich Inc  
United States of America

**Professor. Fadiel Ahmed**  
Tennessee State University  
United States of America

**Professor. Yu Xue**  
Huazhong University of Science and Technology  
China

**Associate Professor Chang-Tsun Li**  
University of Warwick  
United Kingdom

**Professor. Calvin Yu-Chian Chen**  
China Medical university  
Taiwan

### **EDITORIAL BOARD MEMBERS (EBMs)**

---

**Assistant Professor. M. Emre Celebi**  
Louisiana State University  
United States of America

**Dr. Ganesan Pugalenth**  
Genome Institute of Singapore  
Singapore

**Dr. Vijayaraj Nagarajan**  
National Institutes of Health  
United States of America

**Dr. Wichian Sittiprapaporn**  
Mahasarakham University  
Thailand

**Dr. Paola Lecca**  
University of Trento  
Italy

**Associate Professor. Renato Natal Jorge**  
University of Porto  
Portugal

**Assistant Professor. Daniela Iacoviello**  
Sapienza University of Rome  
Italy

**Professor. Christos E. Constantinou**  
Stanford University School of Medicine  
United States of America

**Professor. Fiorella SGALLARI**  
University of Bologna  
Italy

**Professor. George Perry**  
University of Texas at San Antonio  
United States of America

**Assistant Professor. Giuseppe Placidi**  
Università dell'Aquila  
Italy

**Assistant Professor. Sae Hwang**  
University of Illinois  
United States of America

**Associate Professor Quan Wen**  
University of Electronic Science and Technology  
China

**Dr. Paula Moreira**  
University of Coimbra  
Portugal

**Dr. Riadh Hammami**  
Laval University  
Canada

**Dr Antonio Marco**  
University of Manchester  
United Kingdom

**Dr Peng Jiang**  
University of Iowa  
United States of America

**Dr Shunzhou Yu**

General Motors Global R&D Center  
United States of America

**Dr Christopher Taylor**

University of New Orleans  
United States of America

**Dr Horacio Pérez-Sánchez**

University of Murcia  
Spain

## TABLE OF CONTENTS

Volume 8, Issue 1, August 2014

### Pages

1 - 7      Bliss: A New Read Overlap Detection Algorithm  
*Sayali Davalbhakta*



# Bliss: A New Read Overlap Detection Algorithm

**Sayali Davalibhakta**

*Department of Computer Engineering and IT  
College of Engineering, Pune  
Shivaji Nagar, 411005, India*

*sayalid@gmail.com*

---

## Abstract

Many assemblers carry out heuristic based overlap detection to avoid string comparisons. But heuristics skips many true overlaps. Also, the order of the number of read pairs compared for overlaps is higher than the order of  $n$ . In the raw approach it would be  $n^2$  where every read is compared to every other read. Some assemblers have used a hybrid approach to bring the order down from  $n^2$ . Here is an algorithm which works with 100% accuracy. As there is no heuristics involved, it is able to report all the overlaps in a given set of reads without actual string comparisons. It achieves this purely by querying the k-mer position data. Moreover, the number of read pairs compared is proportional to the number of reads present i.e. of the order of  $n$ .

**Keywords:** Overlap Detection, De Novo Assembly, Overlap-layout-consensus, OLC Assemblers, DNA Sequencing, Whole Genome Shotgun Assembly.

---

## 1. INTRODUCTION

As the sequencing techniques advance, the assemblers need to adapt to the trend. The EULER assembler introduced the use of de Bruijn graphs in assembly [1]. These assemblers suffer from heavy memory consumption [2]. The other approach conventionally taken by some de novo assemblers is the overlap-layout-consensus (OLC) approach [3]. The primary difference between the two approaches is that de Bruijn graph is a graph with k-mers as nodes and an edge corresponds to k-1 characters of overlap. Whereas the read overlap graph has reads as nodes and edges correspond to the overlap between reads. Velvet and SOAPdenovo are examples of de Bruijn based assemblers [2], [4]. De Bruijn graph based assemblers suffer from the problem of huge memory consumption. Several solutions have been devised by assemblers to tackle the memory consumption issue. ABySS takes a distributed approach towards building the de Bruijn graph [5], [6]. Some others try representing the de Bruijn graph as a sparse graph [7], [8]. Celera, ARACHNE are examples of OLC assemblers [9].

The OLC assemblers typically have overlap detection among one of the initial phases. The very basic approach would compare every read with every other read for overlap. Thus it will require  $N^2$  comparisons. This has been optimized by assemblers. During the overlap detection phase, ARACHNE generates k-mers and stores along with the source read information [9]. It then sorts the k-mer table so that reads sharing k-mers are adjacent. Only the reads sharing some k-mers are considered for overlap detection. This is great improvement over the raw  $N^2$  comparison approach.

This paper makes two significant advances over this approach to further improve performance. The only overlaps of interest are the ones that occur at the ends i.e. end of one read overlaps with the beginning of another read. Two reads having overlap which does not span till the end are of no use in read alignment. Such overlap is due to repeats in the original sequence. Bliss, the new algorithm proposed in this paper, segregates the k-mers in two different data structures. Assuming the k-mer positions in a read starting at zero, the zero<sup>th</sup> k-mers of all reads i.e. the ones occurring in the beginning of reads, are stored separately along with the source read. There is no

need to store the position separately because it is zero for all of them. The other k-mers at non-zero locations are stored in another data structure along with the source read and the position information. During overlap detection, the data structure of non-beginning k-mers is searched for reads sharing any of the k-mers at zero<sup>th</sup> position. For n reads, at the most, there will be n beginning k-mers. A read is compared with only those reads that share its zero<sup>th</sup> k-mer. Thus the number of read pairs compared is of the order of n.

Bliss detects overlap purely by searching the k-mer position data. The other advancement made by Bliss is by a factor of K. Although k-mers are generated by sliding a window of size K over the reads, during overlap detection there is no need to slide by a position of one. Since one shared k-mer means an overlap of K, the overlap detection can continue by advancing by K positions at each step. Thus it further improves the speed of overlap detection.

At this point of time, error correction is beyond the scope of this algorithm. It operates post error correction. In future when the quality of the sequencing data becomes near accurate, algorithm like Bliss will certainly play a key role in further enhancing the quality of assembly as it is able to trap all the real overlaps unlike the heuristics based approaches.

## **2. OVERLAP DETECTION ALGORITHM**

Bliss takes a three phase approach towards overlap detection.

### **2.1 Phase I – k-mer Generation**

In the first phase, k-mers are generated from each read by sliding a window of size K over the read. The zero<sup>th</sup> k-mers, that is the ones occurring at the beginning of the reads, are put in a data structure Z along with the source read number. The other k-mers are added to another data structure T along with the source read number and the position in the read. K-mers are stored by their unique radix-4 number and the k-mer strings are discarded [10]. This brings in space saving and also substitutes string comparisons by number comparisons.

### **2.2 Phase II – Gather Read Pairs**

In the second phase, read pairs which may possibly overlap are gathered in another data structure P. P holds the id of first read, the index at which overlap begins in the first read and the id of the second read. The overlap, if it is there, always begins at position zero for the second read hence the index is not stored separately for the second read.

Read pairs are gathered by iterating through the set Z. For each k-mer in the set Z, the set T is searched to extract reads that have this k-mer at some position in them. An entry is made in the set P of the id of the read corresponding to this shared k-mer in T, the corresponding position from T and the id of the read corresponding to this k-mer in Z.

If multiple entries of the same read pair are present in P, then they are processed in the ascending order of the position so that the largest overlap is tested for first.

### **2.3 Phase III – Filter out Read Pairs with Discontinuity in Overlap**

For each entry in P, let r1 be the id of the first read i.e. the one that came from the set T, let p be the corresponding position and let r2 be the id of the second read i.e. from Z. Since the k-mer at index p in the first read, matches the zero<sup>th</sup> k-mer in the second read, there is already an overlap of K characters between these reads. This overlap starts at position zero for the second read. We need to check if it continues till the last position in the first read. If not, this read pair is discarded. To check the continuity till the last character, we start at the last k-mer in the first read and the corresponding k-mer in the second read. If these k-mers match, we move backward by K positions in both the reads at each step until there are K or less characters left at the beginning of the second read. These characters are covered by the zero<sup>th</sup> k-mer, hence there is no need to again match them here. So we just report overlap and break. Checking for continuity backwards starting from the end of the first read has a clear advantage of eliminating false overlaps quickly

without requiring any additional k-mer comparisons. This is another speed improvement in the process of overlap detection. The maximum number of k-mers compared i.e. before reporting overlap in this phase is  $\lceil ((l - p)/K) \rceil - 1$  where  $l$  is the length of  $r_1$ . The last k-mer position is computed for the first read by subtracting  $K$  from the read length. If k-mers in both the reads are not identical at any step, there is discontinuity in overlap so we break and move to the next read pair

## 2.4 Algorithm

**Input:**  $R: \{r\}$  Set of reads

$K$ : k-mer length.

**Output:**  $P: \{(r_1, i, r_2)\}$ .  $r_1$  and  $r_2$  overlap from  $i$  to  $(l - 1)$  in  $r_1$  where  $l = \text{length}(r_1)$  and from  $0$  to  $(l - 1 - i)$  in  $r_2$ .

**Data Structures:**

$T$ : Set of 3-tuples  $(k_{mn}, r_m, n)$  k-mer  $k_{mn}$  occurs in read  $r_m$  at  $n^{\text{th}}$  position where  $n \neq 0$

$Z$ : Set of 2-tuples  $(r_m, k_{m0})$   $k_{m0}$  occurs in read  $r_m$  at  $0^{\text{th}}$  position

$P$ : Set of 3-tuples  $(r_1, i_1, r_2)$  Reads  $r_1$  and  $r_2$  share a k-mer at  $i^{\text{th}}$  position in  $r_1$  and  $0^{\text{th}}$  position in  $r_2$ . A possible overlap between  $r_1$  and  $r_2$  may begin here.

### Phase I - Tokenization/K-mer generation

For each  $r_m$  in  $R$ ,

```
{
   $l_m = \text{length}(r_m)$ ;
   $\text{lastKmerIndex} = (l_m - K)$ ;
  for ( $i = 0, i \leq \text{lastKmerIndex}, i++$ )
  {
    /* Get a substring of length K in  $r_m$  starting at  $i$ , Unique number is radix-4 representation of this string */
     $k_{mi} = \text{UniqueNumber}(\text{substring}(r_m, i, K))$ 
    If ( $i \neq 0$ )
    {
      Enter ( $k_{mi}, r_m, i$ ) in  $T$ ;
    }
    Else
    {
      Enter ( $r_m, k_{mi}$ ) in  $Z$ ;
    }
  }
}
```

### Phase II – Gather read pairs with possible overlap occurring at the end

For each  $r_m$  in  $(r_m, k_{m0})$  in  $Z$ ,

- Fetch from  $T$ ,  $(k_{sv}, r_s, v)$  where  $k_{sv} = k_{m0}$  and  $(r_s \neq r_m)$
- Enter in  $P$   $(r_s, v, r_m)$

For multiple entries of  $r_s$  and  $r_m$ ,  $(r_s, x, r_m)$  order by  $x$  ascending so that the largest overlap is tested for first

### Phase III – Filter read pairs with discontinuity in overlap

For each  $(r_s, v, r_m)$  in  $P$ ,

```
{
   $l_s = \text{length}(r_s)$ ;
   $\text{lastKmerIndex} = (l_s - K)$ ;
   $i = \text{lastKmerIndex}$ ;
   $j = (\text{lastKmerIndex} - v)$ ;

  While ( $j > 0$ )
```

```

{
  Search T for tuples  $(k_{si}, r_s, i), (k_{mj}, r_m, j)$  where  $k_{si} = k_{mj}$ 
  If (found)
  {
     $i = i - K;$ 
     $j = j - K;$ 
  }
  Else /* If (found) - Discontinuity in overlap */
  {
    Remove  $(r_s, v, r_m)$  from P;
  }
} /* While */

If  $(j \leq 0)$  /* Overlap detected, remove other entries for this read pair from P */
{
  For all  $q > v$ 
    Remove from P  $(r_s, q, r_m);$ 
}

} /* For Each */

Output P /* P contains all the overlapping read pairs */

```

### 3. DISCUSSIONS

Bliss, by segregating zero<sup>th</sup> k-mers, greatly reduces the number of comparisons. It only considers full overlaps for both the reads. Partial overlaps are partly discarded in the first step by this segregation. The rest of them are removed in the third phase. The number of k-mers looked up for read pair  $(r_s, v, r_m)$  before reporting overlap are  $(\text{ceil}((l - v)/K) - 1)$ , where  $l$  is the length of the first read  $r_s$  and the overlap begins at index  $v$  for the first read.

The data structures need to be indexed to facilitate faster fetch. Suggested indexes are: set T  $\{(k_{mn}, r_m, n)\}$  has two indexes: one on  $k$  and the other on  $(r, n)$ , set P has index on  $(r_1, r_2, i)$  sorted ascending by  $i$  for multiple entries of the same read pair.

Bliss uses unique number representation of k-mers and discards k-mer strings. This further improves speed.

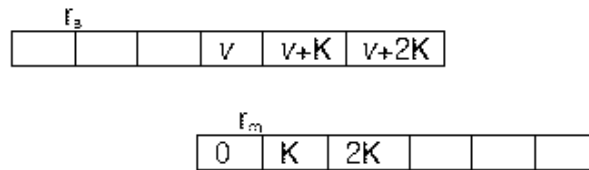
Consider the values given in Table 1.

K-mers	Occurring at position zero in reads - Z	Occurring at non-zero position in reads - T
$k_0$	$r_1, r_{23}$	$r_{13,8}, r_{12,5}, r_{13,22}$
$k_1$	.....	.....
$k_2$	.....	.....

**Table 1:** Example of Shared k-mers.

k-mer  $k_0$  in Table 1 is at zero<sup>th</sup> position in  $r_1, r_{23}$  and non-zero position for  $r_{13, 8}, r_{12, 5}, r_{13, 22}$ . For each row, we compare each read in set Z with each entry in set T. Thus, the pairs compared for k-mer  $k_0$  are  $(r_1, r_{13}), (r_{23}, r_{13})$  starting from 8<sup>th</sup> position in  $r_{13}, (r_1, r_{12}), (r_{23}, r_{12})$  starting from 5<sup>th</sup> position in  $r_{12}$  and  $(r_1, r_{13}), (r_{23}, r_{13})$  starting from 22<sup>nd</sup> position in  $r_{13}$ . Thus there are 6 comparisons done for  $k_0$ .

Figure 1 shows the k-mers compared for  $(r_s, v, r_m)$ . In phase 3 of the algorithm, Bliss iterates for  $r_s$  through  $i = v + 2K, v + K$  and for  $r_m$  through  $j = 2K, K$ .



**FIGURE 1.** k-mers Compared For Detecting Overlap Between Reads  $r_s$  and  $r_m$ .

#### 4. Number of Comparisons

Let  $N$  be the number of reads in the given set,  $L$  be average read length and  $K$  be k-mer length. Thus the total number of k-mers generated in the set is  $(L - K + 1) * N$ . The probability that read  $R_x$  is not compared with read  $R_1$  is  $(1 - 1 / ((L - K + 1) * N))^{(L - K)}$ . Hence the probability that  $R_x$  will be compared with  $R_1$  is  $1 - [(1 - 1 / ((L - K + 1) * N))^{(L - K)}]$ . Thus Expected number of reads that will be compared with  $R_1$  is  $(N - 1) * [1 - [(1 - 1 / ((L - K + 1) * N))^{(L - K)}]]$ . The total expected number of comparisons thus becomes  $N * (N - 1) * [1 - [(1 - 1 / ((L - K + 1) * N))^{(L - K)}]]$ . In this computation, let the total expected number of comparisons be  $N * X$ . The  $X$  above approaches 1 when the real data numbers are put in. Thus the number of comparisons approaches  $N$ .

The test data had 100 ( $N$ ) reads of length 30 ( $L$ ) and  $K$  was 6. The expected number of comparisons becomes: 95. The E.Coli set had 400000 reads with average length of 230. The  $K$  was set to 18. Thus  $N = 400000, L = 230$  and  $K = 18$ . Thus the expected number of read pairs compared becomes: 398121.

In reality, all k-mers are not distinct or else there will be no overlap among reads. Thus the number of distinct k-mers under consideration is less than  $((L - K + 1) * N)$  which slightly increases the number of comparisons and makes it approach closer to  $N$  from the lower side. E.g. consider an ideal case where each read has its last k-mer overlapping with the zero<sup>th</sup> k-mer of exactly one read. The larger is the number of k-mers overlapping, the less is the number of distinct k-mers present. Similarly the more is the number of reads sharing the last k-mer, the less is the number of k-mers. As an example let's take up a case where only the last k-mer overlaps and that too with only one read. This will reduce the number of distinct k-mers by  $N$  making it  $(L - K) * N$ . With these many k-mers in the set, the number of pairs compared for test data and E.Coli above become 99 and 399999 respectively i.e. the numbers are truly linear. This case we considered is the maximum number of k-mers present in the set with overlap present. In reality multiple k-mers will overlap for many read pairs and many k-mers will be present in more than two reads. Thus the number of distinct k-mers will further reduce.

The raw approach would involve  $N * (N - 1)$  comparisons which becomes 9900 and 159999600000 for test data and E.Coli respectively. The approach taken by some assemblers where two reads sharing any k-mers are compared will involve  $N * (N - 1) * [1 - [(1 - 1 / ((L - K + 1) * N))^{(L - K + 1)}]]$  comparisons. For test data and E.Coli, these numbers become 2190 and 85177107 respectively. Clearly Bliss involves the least number of comparisons among these three approaches and yet reports all true overlaps.

## 5. IMPLEMENTATION AND RESULTS

Bliss can be implemented with the data structures T, Z, P as in-memory data structures or persistent data structures. The indexing and data retrieval infrastructure can be implemented from scratch for fine tuning and better optimization or some existing infrastructure can be utilized. I chose persistent data structures. Also, instead of reinventing the wheel, I decided to exploit the indexing and data retrieval facilities provided by a relational data engine to implement the data access layer. To utilize the overlaps detected for contig generation, I supplied it with a basic contig generation algorithm. The contig generation algorithm halts when it cannot extend the contig further or if there are more than one paths available for extension.

I implemented Bliss using MySQL database. The first phase of k-mer generation involves importing the reads from .fa file into the database and then generating k-mers from the reads. The second phase computation is saved by defining a view on the sets T and Z. The third phase iterates over the view to check for continuity of overlap in both the reads. The read pairs for which continuity exists are inserted in the read pairs' table. The contig generation algorithm operates on this read pairs' table to form contigs.

Implementing indexing and data retrieval from scratch to fine-tune to the problem at hand may boost performance compared to relying on a relational data engine. However it trades performance with complexity and hence the development effort.

Bliss was tested on experimental data and on single-end E. Coli data. For E. Coli, on a 2 GB, quad-core machine with 32 bit OS, Bliss took approximately 6 hours to output contigs.

Some of the database level optimizations done are as mentioned here. These MySQL parameters were optimized: `innodb_buffer_pool_size` was set to 1024M, `key_buffer_size` to 64M, `table_open_cache` to 4000, `table_definition_cache` to 4000. To improve performance, minimal required indexes were defined. Inserts, updates, deletes were done in bulk wherever possible. Foreign key constraints were removed. Large tables were partitioned.

Another in-memory optimization done in the third phase was to fetch all the k-mers for the given pair of reads at the given indexes in one shot. The k-mers from one read were bitwise XORed with their counterparts in the other read. The results were then bitwise ORed to get a single number. If this number is zero, it means all the k-mers match and there is true overlap between the read pair else there is no overlap.

## 6. CONCLUSION AND FUTURE SCOPE

Bliss is a simple overlap detection algorithm. It uses a hybrid approach of k-mer generation followed by read overlap detection. It significantly reduces the number of comparisons for overlap detection by segregating zero<sup>th</sup> k-mers from the others. It detects the overlap by searching the k-mer position data. While detecting overlap, it advances in steps of K in backward direction to speed up the process.

Since there is no heuristics involved Bliss offers 100% accuracy. All the overlaps in a given set of reads are correctly reported by Bliss.

The data structures can be in-memory or persistent and may use custom indexing and data retrieval infrastructure. If implemented on a relational data engine, several database level optimizations may be done for performance.

Bliss can be extended to include single bit error correction and multiple bit error trapping by analyzing the sequence of differing k-mers between two reads.

When run post error correction, bliss reports all the overlaps present in the given set of reads.

## 7. ACKNOWLEDGEMENTS

My sincere thanks to Dr. S.D. Bhide and Prof. S. P. Gosavi for their support and valuable feedback.

## 8. REFERENCES

- [1] D. R. Zerbino and E. Birney. "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." (2008) *Genome Research*, 18:821–829
- [2] Jeffrey J. Cook, Craig Zilles. "Characterizing and Optimizing the Memory Footprint of De Novo Short Read DNA Sequence Assembly" April, 2009.
- [3] Flicek P, Birney E. "Sense from sequence reads: methods for alignment and assembly." *Nat Methods*. 2009 Nov;6(11 Suppl):S6-S12.
- [4] Ruiqiang Li, Hongmei Zhu, Jue Ruan, et al. (2009, December ). "De novo assembly of human genomes with massively parallel short read sequencing", *Genome Research*, [Online].Available: <http://genome.cshlp.org/content/early/2009/12/16/gr.097261.109.full.pdf+html>
- [5] Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. "De novo Transcriptome Assembly with ABySS" *Bioinformatics* (2009) 25 (21): 2872-2877.
- [6] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, Inanc Birol. "ABySS: A parallel assembler for short read sequence data", (2009) 19(6):1117-23.
- [7] Chengxi Ye, Zhanshan (Sam) Ma, Charles H. Cannon, Mihai Pop, Douglas W. Yu. SparseAssembler: de novo Assembly with the Sparse de Bruijn Graph  
[Online].Available:  
<http://arxiv.org/ftp/arxiv/papers/1106/1106.2603.pdf>
- [8] Chengxi Ye, Charles H. Cannon, Zhanshan (Sam) Ma, Douglas W. Yu, Mihai Pop. SparseAssembler2: Sparse k-mer Graph for Memory Efficient Genome Assembly.  
[Online].Available:  
<http://arxiv.org/ftp/arxiv/papers/1108/1108.3556.pdf>
- [9] Serafim Batzoglou, David B. Jaffe, Ken Stanley, et al. "ARACHNE: A Whole-Genome Shotgun Assembler" (2002) *Genome Research*, 12:177–189.
- [10] Sayali Davalibhakta. "Throwing Away k-mer Strings" Submitted for publication.

## INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Biometric and Bioinformatics (IJBB)* brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with Volume 9, 2015, IJBB will appear with more focused issues related to biometrics and bioinformatics studies. Besides normal publications, IJBB intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

### LIST OF TOPICS

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- DNA assembly, clustering, and mapping
- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation
- E-health
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks



**CALL FOR PAPERS**

---

**Volume: 9 - Issue: 1**

**i. Paper Submission:** November 30, 2014      **ii. Author Notification:** December 31, 2014

**iii. Issue Publication:** January 2015

## **CONTACT INFORMATION**

### **Computer Science Journals Sdn Bhd**

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6204 5627

Fax: 006 03 6204 5628

Email: [cscpress@cscjournals.org](mailto:cscpress@cscjournals.org)

CSC PUBLISHERS © 2014

COMPUTER SCIENCE JOURNALS SDN BHD

B-5-8 PLAZA MONT KIARA

MONT KIARA

50480, KUALA LUMPUR

MALAYSIA

PHONE: 006 03 6204 5627

FAX: 006 03 6204 5628

EMAIL: [cscpress@cscjournals.org](mailto:cscpress@cscjournals.org)