deSPI: efficient classification of metagenomics reads with lightweight de Bruijn graph-based reference indexing

Dengfeng Guan & Bo Liu & Yadong Wang

Dept. of Computer Science and Technology, Harbin Institute of Technology

dfguan@hit.edu.cn

Abstract

One of the core problems in metagenomics is the classification of shotgun sequencing reads to identify species present in samples. Many supervised classification tools have been developed recently, but they either consume large memory or large computation time. Herein we propose a new classification method, de Bruijn Graph-based Species Identifier (deSPI), which takes advantage of de Bruijn graph and FM-index data structures and a hierarchical top-down strategy to do classification. The experimental results suggest that deSPI uses much less memory than Clark and Kraken and classifies reads much faster than Centrifuge and Kaiju, while maintaining a comparable sensitivity and accuracy.

Souce code: deSPI is available at https://github.com/dfguan/deSPI.

Introduction

Metagenomics has become a major technique to study the microbiome over the last decades. It can help to uncover the complexity of a microbial community in a specific environment, analyze correlation between the community with the environment, and discover new microbial species that cannot be cultivated in the laboratory. With the development of next-generation sequencing (NGS) technologies and decrease of sequencing cost, ever larger metagenomics sequencing data sets are being produced. While these give us more opportunities to perform microbiome analysis, they also generate many analytical challenges. One of the major computational challenges is metagenomics classification.

In recent years, numerious supervised metagenomics classification tools are developed. Many of them are based on short exact matches. Among these tools, Kraken [6], Clark [5], Centrifuge [2], Kaiju [4], are now being widely used. However they either consume a large amount of memory or take too much time to classify. Herein, we propose deSPI, a novel short exact match based metagenomic read classification tool, which has higher speed and affordable memory cost with higher or comparable sensitivity and accuracy.

Methods

deSPI infers taxa of reads through a FM-index and hierarchical top-down taxonomic classification. In the indexing phase (Figure 1), deSPI takes a reference database and its corresponding phylogenetic tree as inputs. Instead of indexing original genomic sequences, it builds a FM-index of unitigs from the de Bruijn graph of references (RdBG) for both strands [3], and each unitig is labeled with a taxon. In the classification phase (Figure 2), it applies a sparse seeding strategy to find maximal exact matches between reads and unitigs (U-MEMs) and infer taxa of the reads with a hierarchical top-down taxonomic classification strategy.

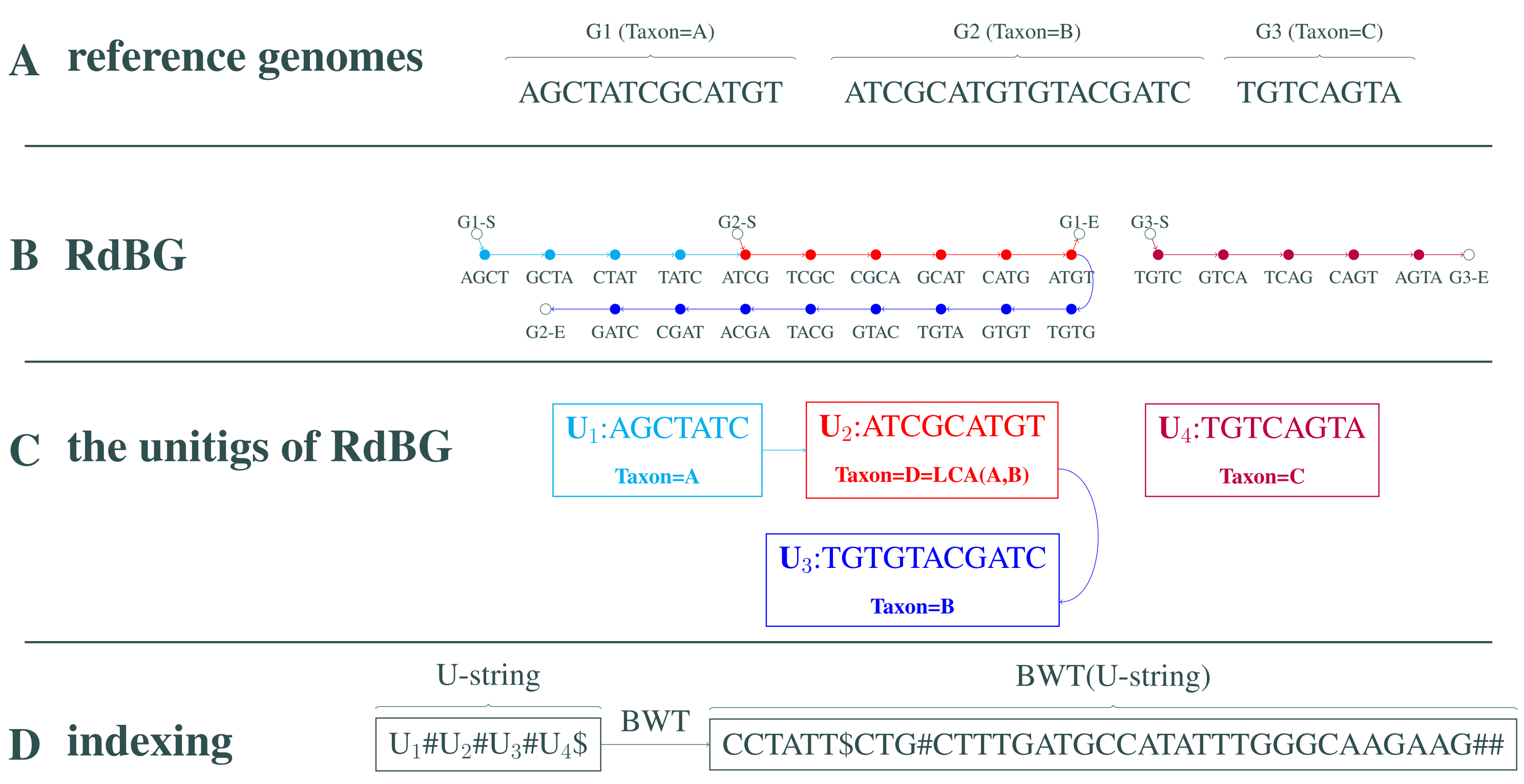
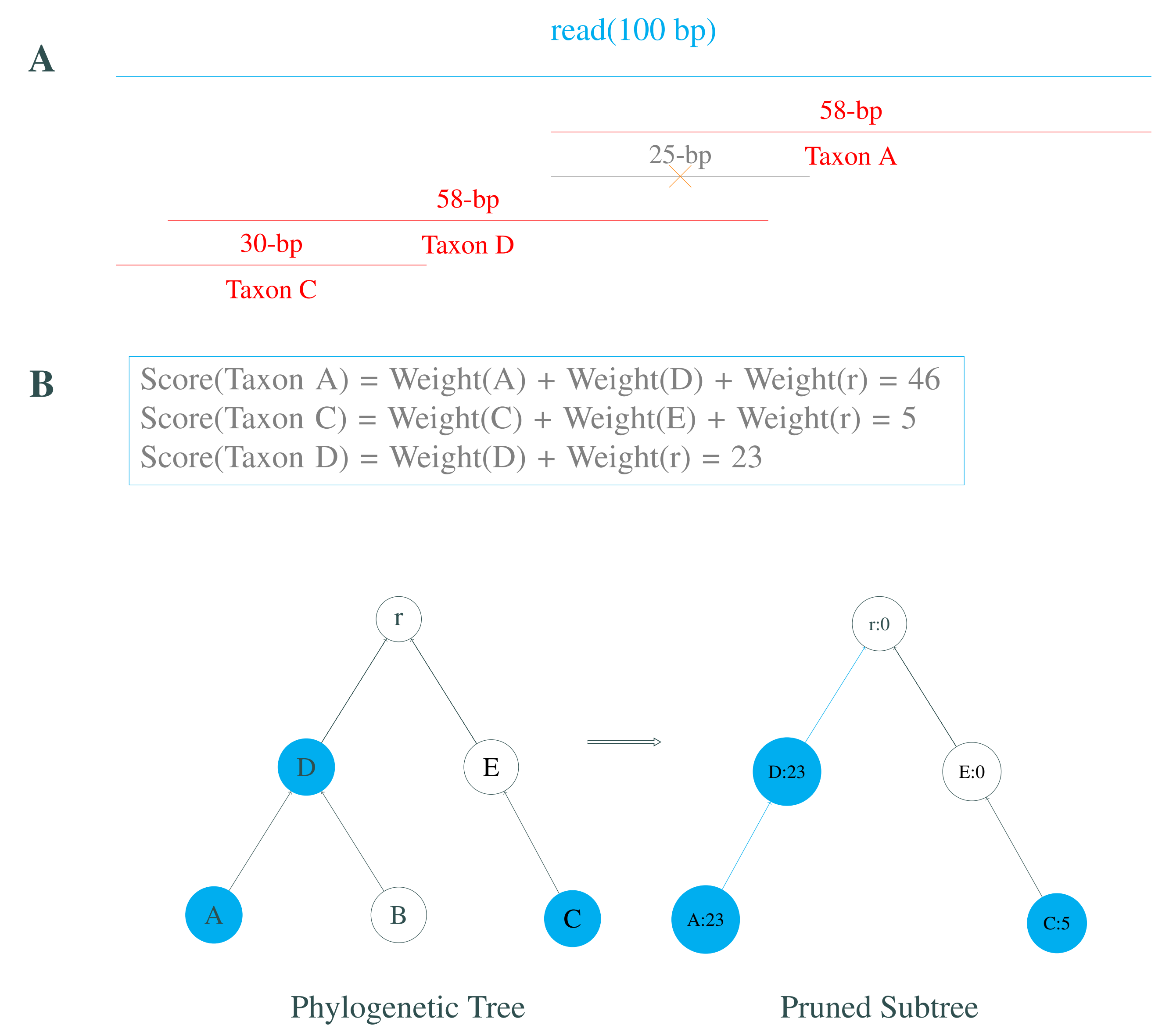


Figure 1: Index Reference Genomes. (**A**) A library of reference genomes. (**B**) Build de Bruijn Graph (RdBG) of the reference database. (**C**) Obtain unitigs from RdBG, each unitig is labeled with a LCA of the organisms containing the unitig. (**D**) Obtain U-string through joining all unitigs together with a delimiter #, perform Burrow-Wheeler Transformation on U-string and generate FM-index.



Score(Taxon A) = Weight(A) + Weight(D) + Weight(r) = 46
Score(Taxon C) = Weight(C) + Weight(E) + Weight(r) = 5
Score(Taxon D) = Weight(D) + Weight(r) = 23

Figure 2: Classification of a read. The figure illustrates the deSPI classification algorithm using seed length 25. (**A**) deSPI performs backward search along the read, find all U-MEMs larger than the minimum seed length. (**B**) Nodes receiving hits (highlighted in cyan) form a pruned phylogenetic tree, deSPI traverses the tree to figure out the classification path (nodes and edges highlighted in cyan).

Results

Three experiments were carried out on simulated and real datasets, to assess the feasibility of deSPI. Four state-of-the-art metagenomics classification tools, Centrifuge, Clark, Kaiju and Kraken are tested on the same datasets for comparison. The *k-mer* used for indexing of Clark and Kraken, and RdBG of deSPI is 31. deSPI was tested with seed lengths of 24, 28 and 31, while the other tools were ran with their default settings. For simulated data and real sequencing data from single individuals, the tests were run on a server with 1TB of Random Access Memory (RAM), eight 2.0 GHZ Intel Xeon E7-4820 CPUs (64 cores). For the real metagenomic data, the test was performed on a server with 767.8 GB of RAM, fifty-six 2.6 GHZ Intel Xeon E5-2690 CPUs (784 cores).

We compared the classification accuracy of deSPI with the other four methods at both species and genus levels. Here we defined rank level sensitivity (Sen in Table 1, 2, 3) as A/B, where A is the number of reads with a correct assigned taxon at the measured rank and B is the number of reads to be classified, accuracy (Acc in Table 1, 2, 3) as $A/(C + D)$, where C is the number of reads with assigned taxon at or below the measured rank, D is the number of reads with an incorrect assigned taxon above the measured rank, and F1-score (F1-score in Table 1, 2, 3) as harmonic mean of sensitivity and accuracy.

		Species	Genus	Memory	Speed
		Sen/Acc/F1-score	Sen/Acc/F1-score	GB	Kseq/m
deSPI	24mer	**76.59**/96.83/**85.53**	**94.28**/99.23/**96.69**	25	822
	28mer	73.64/98.36/84.22	91.45/99.54/95.32	25	853
	31mer	70.68/**99.37**/82.61	88.17/**99.81**/93.63	25	755
Centrifuge		75.06/98.96/85.37	91.86/99.71/95.62	**10**	416
Clark	31mer	70.54/99.27/82.47	70.84/99.69/82.82	79	**1077**
Kaiju		40.23/95.56/56.62	63.38/98.17/77.03	14	195
Kraken	31mer	71.83/99.26/83.34	89.39/99.78/94.30	126	722

Table 1: Classification performance for deSPI, Centrifuge, Clark, Kaiju and Kraken on 20 million simulated reads

			HiSeq Single-end			MiSeq Single-end		
		Species	Genus	Speed	Species	Genus	Speed	Memory
		Sen/Acc/F1-score	Sen/Acc/F1-score	Kseq/m	Sen/Acc/F1-score	Sen/Acc/F1-score	Kseq/m	GB
deSPI	24mer	**34.46**/94.98/**50.57**	**49.05**/97.55/**65.28**	1562	**21.45**/68.76/**32.70**	**56.36**/95.86/**70.99**	1004	25
	28mer	34.11/96.70/50.43	47.91/98.36/64.43	1600	21.05/72.25/32.60	55.33/96.92/70.45	993	25
	31mer	33.80/**97.44**/50.19	46.86/**98.69**/63.55	1580	20.79/74.19/32.48	54.64/**97.36**/70.00	969	25
Centrifuge		34.20/93.72/50.11	48.21/97.89/64.60	715	20.85/68.81/32.00	55.00/96.29/70.01	453	**10**
Clark	31mer	33.74/97.20/50.09	34.02/98.02/50.51	**2060**	20.74/**74.56**/32.45	26.43/95.01/41.35	**1074**	79
Kaiju		19.09/91.18/31.57	30.11/95.44/45.78	262	9.21/62.14/16.04	42.85/94.18/58.90	231	14
Kraken	31mer	33.89/97.32/50.27	47.29/98.66/63.93	1143	20.89/73.70/32.55	54.92/97.27/70.20	873	126

Table 2: Classification performance of all tools on 35 single-end datasets

			HiSeq Paired-End			MiSeq Paired-End		
		Species	Genus	Speed	Species	Genus	Speed	Memory
		Sen/Acc/F1-score	Sen/Acc/F1-score	Kseq/m	Sen/Acc/F1-score	Sen/Acc/F1-score	Kseq/m	GB
deSPI	24mer	**35.04**/52.00/41.87	**69.29**/88.76/**77.83**	537	**74.10**/82.20/77.94	**89.94**/96.13/**92.93**	246	25
	28mer	34.43/55.16/42.39	67.68/91.37/77.76	544	73.73/83.77/78.43	89.08/97.12/92.93	230	25
	31mer	33.94/**56.76**/42.48	66.50/**92.34**/77.32	526	73.39/**84.59**/**78.59**	88.44/**97.53**/92.76	201	25
Centrifuge		34.39/54.19/42.08	67.89/89.65/77.27	365	72.77/83.07/77.58	87.86/96.48/91.97	107	**10**
Clark	31mer	33.90/57.23/**42.58**	53.63/90.57/67.37	**689**	73.34/84.40/78.48	84.46/97.21/90.39	**298**	79
Kaiju		21.69/47.68/29.82	55.64/85.83/67.51	175	32.21/72.08/44.53	50.80/93.51/65.84	105	14
Kraken	31mer	34.09/56.24/42.45	66.92/91.96/77.47	454	73.44/84.35/78.52	88.63/97.42/92.82	213	126

Table 3: Classification performance of all tools on 519 paired-end datasets

		Species	Memory	Speed
		Sen/Acc/F1-score	GB	KSeq/m
deSPI	24mer	34.16/76.91/**47.31**	25	1025
Centrifuge		34.51/72.10/46.68	**9**	971
Clark	31mer	24.57/**96.57**/39.17	70	1630
Kaiju		**48.97**/35.86/41.40	14	319
Kraken	31mer	27.93/86.75/42.25	126	**1673**

Table 4: Classification performance of all tools on 11 real metagenomics datasets. We use classification result of Megablast[1] as a pseudo grand truth set. We defined A as the number of correctly classified reads, B as the number of total reads, C as the number of classified reads. And we calculate accuracy as A/C, sensitivity as C/B, and F1-score as harmonic mean of sensitivity and accuracy.

Conclusions

Herein, we propose a novel algorithm, deSPI, for the classification of metagenomics reads. deSPI indexes unitigs of a de Bruijn graph of a reference database with FM-index, and performs top-down hierarchical taxonomic classification to infer the reads' label. One of the advantages of deSPI is that, with its innovative genome indexing approach, it can use several folds smaller RAM space than *k-mer* index based approaches like Kraken and Clark, to achieve similar speed, which is faster than other proposed approaches using small size index, e.g., Centrifuge and Kaiju. As the memory footprint is only around 25 GB when using all RefSeq complete genome sequences as reference, deSPI can fit to the configurations of most modern PCs and workstations. Moreover, the classification results on hundreds of simulated and real sequencing datasets demonstrated that, in most cases, deSPI can achieve overall the best classification results (F1-score), and its accuracy is also comparable to (if not higher than) state-of-the-art approaches. Considering its lightweight, fast speed and relatively good classification ability, we believe deSPI can have enormous potential to be applied in metagenomics data analysis.

We will focus on two studies in the future. One is to develop a fast approach to build the FM-index of unitigs for a large reference database, e.g., RefSeq database. Currently it takes almost 5 days, 500 Gigabytes for deSPI to build the index for complete RefSeq bacterial, archaeal and viral genomes with default settings on a server with 1TB RAM, eight 2.0 GHZ Intel Xeon E7-4820 CPUs (64 cores), it is computationally expensive. The other one is to further improve the sensitivity and accuracy of deSPI.

Acknowledgements

This work has been supported by the National Key Research and Development Program of China (No. 2018YFC0910504 and No. 2017YFC0907500), and China Scholarship Council. And we thank Richard Durbin for comments on the manuscript.

References

[1] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, October 1990.

[2] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, 26(12):1721–1729, December 2016.

[3] Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. deBGA: read alignment with de bruijn graph-based seed and extension. *Bioinformatics*, 32(21):3224–3232, November 2016.

[4] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nat. Commun.*, 7:11257, April 2016.

[5] Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16:236, March 2015.

[6] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, March 2014.