Google Summer of Code 2021

Chapel

Project - Bioinformatics Benchmarks

Introduction

I am Shubham Kumar, an Engineering physics undergraduate student at **Indian Institute of Technology Roorkee.** I am currently in my third year of graduation. I started learning programming languages when I was in 12th grade, the first-ever programming language I learned was C++, since then I have been exploring different concepts of Computer science. I love learning different data structures and algorithms. I also have good experience in the field of website development, with a handful of experience in data science, machine learning, and ethical hacking.

Relevant course work for the project

- Data structures
- Algorithms
- Parallel programming
- Performance Analysis

Motivation for the project

Why do you wish to participate in the Google Summer of Code?

I have been a software enthusiast since when I first started learning about programming. I learned about website development in my first year of college. That's how I got introduced to git and GitHub. I did some web-dev-based projects with my college friends using Github that ultimately introduce me to open-source. For me, It would be satisfying and completely worth it to contribute to a big open-source organization because my work would be helping a lot who uses that open-source software and by participating in the GSoC, I would get to know more about open-source. My development skills would be enhanced and I would be able to build

a strong network with other contributors all over the world, with the guidance of mentors who have a lot of experience in this field. With the remote working policy of GSoC, it is the best program a college student can get involved in, to upgrade his skills and learn a new set of skills during the summer holidays.

Why do you wish to work with the Chapel project in particular?

The chapel project is a great opportunity for me to dive into the field of parallel programming. I always loved the subject biology and topics related to human genome sequences when I was in school and the project I am willing to work with Chapel is to compare an algorithm on genome sequences with other languages so I am keen to apply the algorithm with the help of Chapel's parallel programming features.

I have been learning about Chapel for the last 1-2 months and I have done some contributions also. The best thing about the Chapel community is that they explain the issues very well so a beginner can also easily learn the language. The mentors of Chapel are very helpful and responsive, they are always ready to help us in understanding the concept to any extent.

While I was trying my hands on some issues on Chapel, members of the Chapel have helped me a lot. I came to know about so many new concepts of Chapel after interacting with them. I would like to thank all of the members of Chapel for guiding me through the process of understanding the Chapel.

What do you hope to learn over the summer?

Over the summers I would be learning about K-mer counting algorithms that are used in bioinformatics and metagenomic analysis. K-mer counting implementation in Chapel would be a better scale to the large problem sizes as Chapel provides a high-level interface to distributed parallelism and shared-memory parallelism. Working on the project in the summers would help me to upgrade my skill sets with the help of experienced members and contributors of Chapel.

How well can you comprehend and understand English? How strong is your written English?

I live in India so my first language is Hindi but I am proficient in reading, writing, and understanding English as well. Most of us here in India are taught the English language from childhood. I also had an English communication course in my first semester of college so I am pretty comfortable with the language.

Do you have any other commitments for the summer period? Do you have planned vacations?

My semester end-term examinations will be held in the middle of May so my summer vacations are starting from around 1st June and ending on 30th July, and GSoC's official period is from 7th June to 23rd August. I can easily give my 35-40 hours per week during the vacations and once my college starts, I can still manage 30-35 hours per week. Other than that, I have neither any pre-planned commitments for the summer period nor any planned vacations, so I can give most of my time to GSoC.

Contact

Name	Shubham Kumar
University	Indian Institute of Technology Roorkee
GitHub	shubhamkmr04
E-mail	shub.clares2015@gmail.com
Location	Roorkee, Uttarakhand, India
Timezone	IST(UTC + 5:30)

The timezone does not change during summertime. The time I would be comfortable working in:

- UTC 04:00 - 07:30 hours (ITC 09:30 - 13:00 hours)
- UTC 09:00 - 13:30 hours (ITC 14:30 - 19:00 hours)
- UTC 16:30 - 19:30 hours (ITC 22:00 - 01:00 hours)

The time I have given is not rigid, I can even change my schedule if my currently given schedule is having conflicts communicating with other developers.

Coding experience

Languages: Chapel, C/C++, Python, JavaScript, SQL, Matlab, HTML, CSS
Database: MongoDB

Other technologies I know:

ReactJS- An open-source, frontend, javascript library for building user interfaces.
NodeJS- An open-source, backend, javascript runtime environment that executes javascript code outside a web browser.
ExpressJS- A web application framework for NodeJS.

My experience with different programming languages is shown below in the table:

Language/Library/Framework	Experience (In years)	Level (1-10)
C/C++	3.5	9
Python	1.5	8
Javascript	1.5	7
Matlab	0.5	5
HTML	2	10
CSS	2	9
SQL	1	6
ReactJS	1.5	7
NodeJS	2	8
ExpressJS	2	8

Chapel

I started learning Chapel around 1-2 months ago. This is what my Chapel learning process looks like :

- I watched the youtube video Chapel Comes of Age by Brad Chamberlain that's how I got introduced to Chapel.
- I went through Learn X in Y minutes docs for Chapel and tried some examples.
- I read the official documentation for the Chapel which introduced me to so many features of Chapel.
- I made some pull requests for Chapel on Github.

After some research and self-learning, I am pretty comfortable in writing and understanding the code in Chapel.

Are you familiar with the following tools?

- **git/make-** I have been using these tools with my college development team so I am quite comfortable using them.
- **gnu/Valgrind-** I have used these debuggers a couple of times in my first year of college but right now, I am not that comfortable with them, I would learn them.
- **gcc-** I use the gcc compiler for compiling the C++ code, so I am quite familiar with this.

Experience

I have been diving into the world of programming for 2-3 years now. In this time, I got the chance to meet so many experienced developers. My main developing experience is as follows:

Web development, Thomso (Member):

I was an active member of the web development team of **Thomso**, one of the largest cultural festivals in India. In the group, the goal of me and my teammates was to handle the official website for Thomso. I learned to work in a disciplined manner as we had to manage over 10,000 guest students coming from all over India to be a part of the festival. Being a part of the team, I worked with many experienced

developers, and I have learned how to work as a team to produce the maximum possible results that would be beneficial to the team.

Ingress, IIT Roorkee:

I have been a part of an Ethical hacking group of my college, named Ingress. I met some experienced software enthusiasts there. The tools I have used while diving through the concepts of ethical hacking are:

- Netsparker
- Kali Linux
- NMAP (Network Mapper)
- Metasploit
- SuperScan

Microsoft Codefundo (2019):

My team's ideas got selected for Microsoft Codefundo 2019 and we successfully submitted our azure blockchain app for online voting.
We deployed an online voting app that allows the user to create new accounts and change their information online with the safety of blockchain.
Here is the code we have written: Azure Blockchain App

Open-Source

Is any of the code you have written already open-source? Can you point us to the code you have written?

Yes, I have written an application called "ShubConnector" which is basically like a social media for developers. We can create accounts with a password, make profiles, and create posts to connect with the other developers.
The code is available here on my GitHub: ShubConnector

Contributions to Chapel:

I have been learning about Chapel for the last 1-2 months now. While going through the code base, I have learned so many new and specific features about this Chapel that I have not seen in any language before. I tried to make some pull requests and solve some issues which are being listed below-

Pull requests	Issue (If any linked)	Description	Status
#17400	#17385	Fixed Man Page bug with double-dash arguments	**Merged**
#17446	--	Fixed a Comment in doc/rst/conf.py	**Merged**
#17455	--	Locally overridden Manpage configuration	**Open**
#17497	#17469	Updated documentation for Error.message() to reflect string return type	**Merged**
#17516(Closed) #17555	#17506	Added commutativity in string/bytes multiplication	**Open**

Survey

Have you heard about Chapel before the Summer of Code? If so, where? If no, where would you advise us to advertise?

No, unfortunately, I did not hear about Chapel before Summer of Code. I just came to know about Chapel 2 months ago when I was looking through different open source organizations to contribute but Chapel caught my eye in the first place as I was keen to learn something new and Chapel was a whole new language that's how I decided to contribute to Chapel for Google Summer of Code.
As a student, I would suggest that the biggest way to advertise Chapel is to organize conferences and webinars in different universities and schools, make the online courses for Chapel so students can learn it online by registering through different online course selling websites.

What will keep you actively engaged with the Chapel community after this summer is over?

I would keep engaging myself through the Chapel community even after this summer is over. Great guidance from Chapel community members would keep me engaged. The best thing I love about the Chapel community is that they always explain the smallest possible doubts in the briefest possible way, that's how my learning graph has peaked from last month. I will always devote myself to Chapel to make it a better and better platform than before.

Are you applying to any other organizations for this year's Google Summer of Code? If so, what is the order of your preference, in case you are accepted to multiple organizations?

No, I don't have any plans to apply to any other organization for this year's Google Summer of Code. I am only applying to Chapel.

Prerequisites

What operating system do you work with?

I work with Ubuntu 18.04 and Windows 10.

Are you able to install software on the computer you plan to use?

Yes, I can install every software I planned to use on my computer.

Will you have access to the computer with an internet connection for your development?

Yes, I will always have an active internet connection on my computer during the development period.

Self-assessment

What does useful criticism look like from your point of view as a committing student?

As a committing student, I think that we should not take the criticism personally, rather we should take it constructively and use it to improve ourselves.

What techniques do you use to give constructive advice? How do you best like to receive constructive feedback?

Personally, If I want to give someone advice, I would first explain the scenario of the current situation and how it can be better by doing it in the other way so they would take the feedback constructively and try to improve themselves. I like to receive constructive criticism, It helps me to improve myself and have healthy communication with others.

What is your development style? Do you prefer to figure out/discuss changes before you start coding? Or do you prefer to code a proof-of-concept to see how it turns out?

I prefer to code a proof-of-concept, then get reviewed by mentors and improve the code as suggested. This technique always helps me to learn from my own mistakes and make my code better.

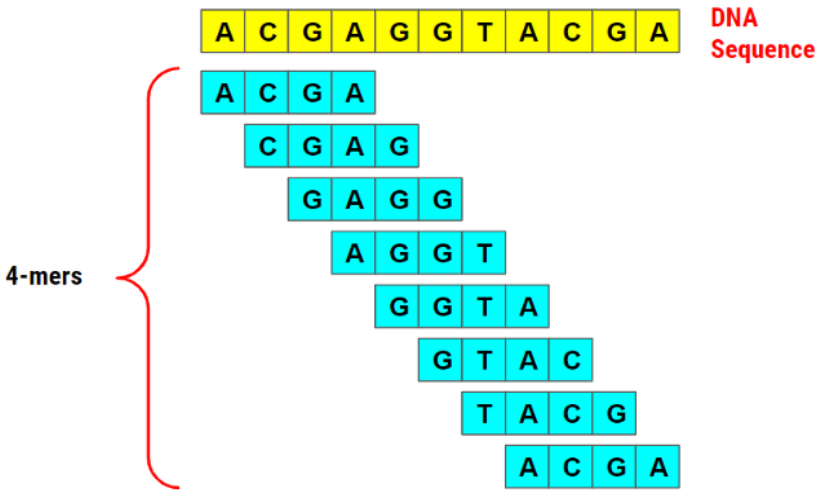
The task

I am working on one of the ideas given in the ideas list, **Bioinformatics Benchmark.** Our task for the project is to write a kmer counting benchmark in Chapel and compare it with existing kmer counting tools in other languages.

What is a k-mer?

A k-mer is just a sequence of k characters in a string (can be DNA, RNA, protein, or any other sequence). To calculate all the k-mers from a string, we need to get the first k characters, then move a single character to get the starting of the next k-mer and so-on.

For example, let's take a DNA sequence "ACGAGGTACGA" containing 11 nucleotides. Let's take k=4 and try to make k-mers from the given sequence.



If we have a sequence of length N, then the formula for finding total k-mers from the given sequence would be N - k + 1 where k is the length of the substring.

Why are k-mers so popular?

K-mers are used in sequence matching. If we already know the genome of organism A and organism B and we want to know if a sequence comes from organism A or organism B, we can check if a sequence contains more k-mers from organism A or organism B.

Reverse complements and Canonical k-mers :

Reverse complement of a genome sequence is when we reverse the sequence, exchanging “A” to “T”, “T” to “A” and “C” to “G”, “G” to “C”. For example, the reverse complement of **ATCGAC** would be **GTCGAT**.
The canonical sequence is the one that is lexicographically smaller of two reverse complements. Most of the k-mer counting tools use a canonical sequence as default. For example, GTCGAT will be counted as ATCGAC.
Let’s take 3-mers of ATCGATCAC.

Offset	0	1	2	3	4	5	6
3-mer	ATC	TCG	CGA	GAT	ATC	TCA	CAC
Reverse Complement	GAT	CGA	TCG	ATC	GAT	TGA	GTG
Canonical	ATC	CGA	CGA	ATC	ATC	TCA	CAC

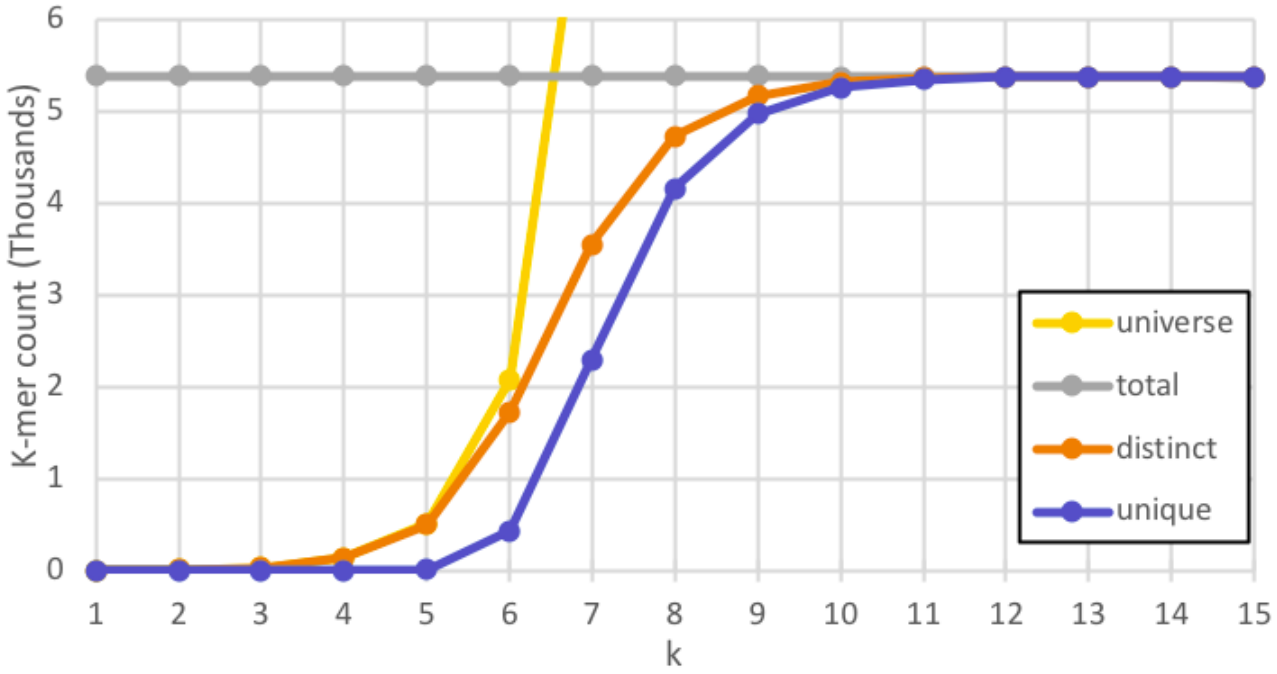
We include both canonical and reverse complements. We get some cases where Canonical form is the original sequence and in other cases, it was a reverse complement.

Unique and Distinct k-mers:

- Distinct k-mers are counted only once even if they appear multiple times.
- Unique k-mers appeared only once.

We call all possible k-mers universe and their count is the size of the universe for a specific value of k. Each nucleotide in a k-mer can be any of the {A, C, G, T}, so the possible combinations for k positions can be 4^k.
Let’s see the graph between the value of k v/s k-mer count for the Phi-X genome.

Phi-X genome



For the given example, we can't consider taking the value of k less than 6, as the number of unique k-mers is almost zero. For k>10, we got unique k-mers.

K-mer counting as Computer's perspective:

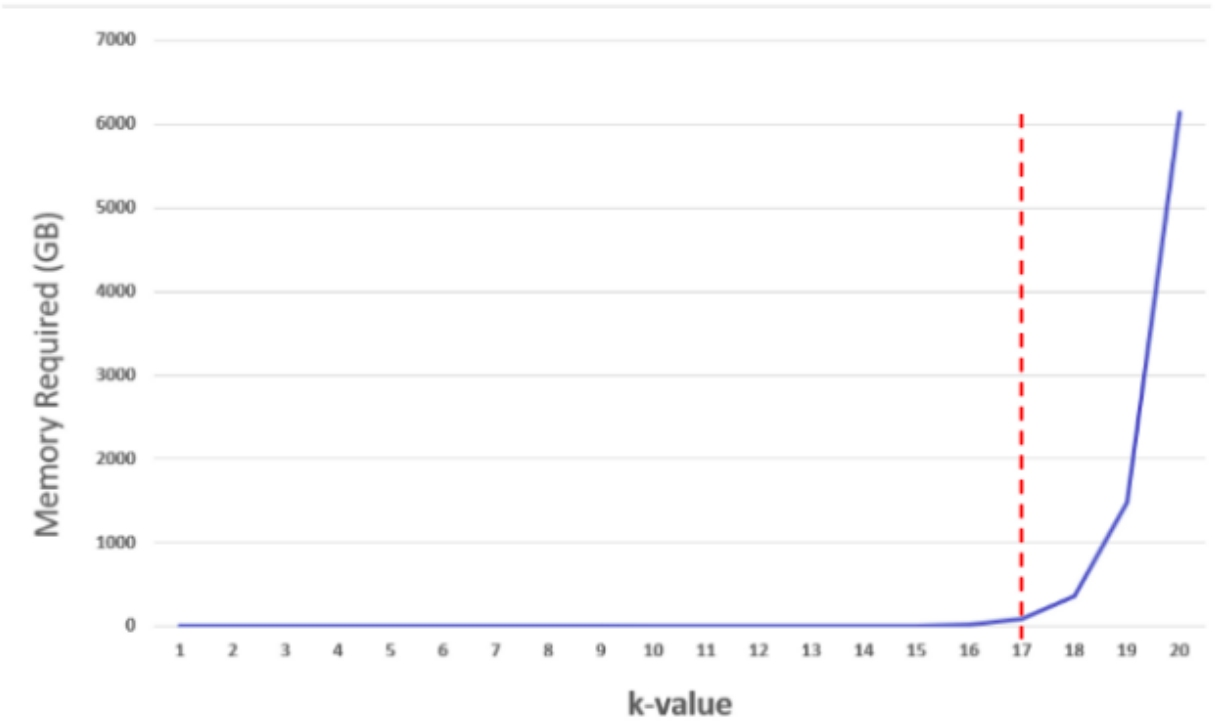
Let's suppose k=3,

Key	Value
AAA	XX

- No. of bits per character = 2
- No. of bits per count = 8

For key, 3*2 =6 bits
For value = 8 bits
Total bytes for 4^3 rows = 14*(8+6) =bytes (896 bits)

As we keep increasing the value of k, we need more memory as shown in the following graph:



If we keep k=20, we need almost 6000 GB of RAM, which is a huge number. This much memory consumption of k-mer is what makes it very challenging for computer scientists.
This is why the Chapel would be better as we can use its task parallelism and shared-memory parallelism to satisfy the heavy memory demands for k-mer countings.

K-mer in Chapel

Here is a simple example of k-mer counting in Chapel using Map Standard Library Module:

```
use Map;

var kmer= new map(string, int);
proc kmercounting(k, sequence){
    for i in 0..(sequence.size - k){
        if(!kmer.contains(sequence[i..#k])){
            kmer.add((sequence[i..#k]) , 1);
        }
        else{
            kmer[sequence[i..#k]]+=1;
        }
```

```
    }
return kmer;
}
var sequence = "GCATCGACGTACGGCAT";
var k:int =4;
writeln(kmercounting(k, sequence));
var distinct:int =0;
var unique:int =0;
for i in kmer.values() {
    if(i==1){
        unique+=1;
    }
    distinct+=1;
}
writeln("Unique kmers: ", unique);
writeln("Distinct kmers: ", distinct);
```

We got the output as we expected:

```
{GTAC: 1, GGCA: 1, ACGT: 1, CGTA: 1, CGAC: 1, CGGC: 1, GCAT: 2,
TCGA: 1, TACG: 1, ACGG: 1, ATCG: 1, GACG: 1, CATC: 1}
Unique kmers: 12
Distinct kmers: 13
```

Motivation

Why is this particular task exciting for me?

This task is exciting for me because it comes under the topic related to biology. I always had a great interest in biology, human DNA, genome sequences, and all these kinds of stuff. I always wanted to do something different on these topics.

What do you hope to learn by working on it?

If I get the opportunity to work on this project, I would be learning a lot of skills by the end of this summer. I would get to learn more aspects of Chapel and how to set benchmarks in programming languages. I am hoping to learn from the amazing developers I would be working with within the summer.

Timeline

17 May 2021 to 6 June 2021	● Get more involved in the Chapel community ● Continue contributing to Chapel ● Would try to solve my opened PR's
7 June 2021 to 16 June 2021	● Implement the k-mer counting parallelly ● Create graphs for performance in different k values
17 June 2021 to 25 June 2021	● Would try k-mer counting in Jellyfish ● Performance analysis with k-mer counting in Chapel if done parallelly
25 June 2021 to 5 July 2021	● Would try memory-efficient data structures like Bloom filters ● Performance analysis for this method ● Graphs for different k values to find out how many unique k-mers we get in
5 July 2021 2021 to 15 July 2021	● Would find more k-mer counting tools in other languages ● Compare them with different k-mer counting approaches in Chapel
16 July 2021 to 15 August 2021	● Would be working on the project under the guidance of mentors ● Would take the help of mentors and try to resolve all my doubts regarding the project to make my project results better

References

- https://bioinfologics.github.io/post/2018/09/17/k-mer-counting-part-i-introduction/
- https://chapel-lang.org/docs/master/modules/standard/Map.html
- https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29
- https://www.youtube.com/watch?v=ANBInLeKHEI&ab_channel=RobEdwards
- https://medium.com/@konnks666/what-do-you-mean-by-performance-analysis-3fa6ef2cc9bb#:~:text=Performance%20analysis%20is%20the%20technique,they%20allotted%20him%20or%20her.