What is GSoC all about?

Plus, my project, experience and some tips.

JaydevSR

October 30, 2022

Dept. of EECS, IISER Bhopal

What is Google Summer of Code?

Some quick pointers:

- Google Summer of Code or GSoC for short is an annual mentorship program.
- This takes place over three months[1] starting in the summer (which was June to September this year).
- It is managed by a team from Google who provides supervision as well as a stipend for the contributors.
- Several organization take part in this program looking for good contributors for their open-source libraries.

We will see more details of the program later.

[1] although the duration can be extended on request

Where do I come in all this?

I was selected for the GSoC 2022 by The Julia Languange for contributing to their molecular dynamics library Molly.jl.

My project included implementing something called *parallel replica exchange algorithms*. This is helpful when biologists want to simulate complex biomolecules such as proteins.

For this I was mentored by Dr. Joe Greener, a molecular biologist by trade, who maintains (and built) this library.

Why you should apply as well?

There are many things that one can achieve by taking part in GSoC, but just to list a few:

1. Makes you learn to write programs that are *efficient* and that others (and your future self) can *understand*!
2. *Get mentored* by people who built the libraries you use. And leave a part of yourself in the libraries you use.
3. A successful GSoC is *undoubtedly a major highlight* in a student's resume.
4. Throughout and after the program you have the opportunity to make a number of *connections* with experts at prominent organizations.
5. Also, this is basically you getting paid for becoming a better programmer yourself.

What is the application procedure?

The overall timeline of the application window is something of this manner:

1. $\sim$ First week of March: List of selected organizations released.
2. $\sim$ First week of April: Contributor application begin (usually open for 2-3 weeks)

The one month period between (1) and (2) is where you are supposed to make initial contacts and finalize projects.

But ideally this should not be your strategy if you want to have best chances of selection!

Most of the organizations are same every year! And many of their projects that are floated remain unclaimed!

Tip

Browse through the past year's organizations and their projects in the ideas list and see what projects were not selected for that year. Those are then up for grabs for you and you can start preparing for those with a headstart.

Surely this is not the only way for you to decide the projects, you can contact the mentors and propose you own project about:

- Something you feel like missing in the library
- Some cool new technique that was recently developed and will be a good addition.
- Some feature or problem that is highlighted in the GitHub issues or the library's webpage.

Examples of some organizations


julia


R

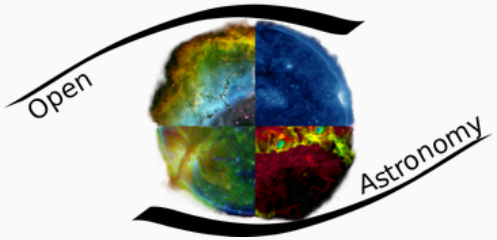

OpenCV


NUMFOCUS
OPEN CODE = BETTER SCIENCE


django

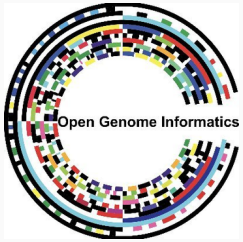

blender


CERN
HSF
HEP Software Foundation


Open
Astronomy


Open
Chemistry


OPEN
BIOINFORMATICS
FOUNDATION


Open Genome Informatics


sage

Choose an organization you care about!

The next thing to do is to reach out!

Once you have an organization(s) in mind that you like, reach out to them immediately, the earlier the better. You can find names of prospective mentors in the ideas list on the organizations page in the GSoC website. The possible ways you can contact your prospective mentors or Org admins are:

- Contact info from their *GitHub profile/webpage*.
- Through the organization's *Slack/Zulip* etc.

Tip

Once you have a project(s) in mind, don't wait and start contributing immediately. This will increase the chance of you selection by integer factors (just a guess!). You can open a draft PR and start adding code. Or, if you are not familiar with all this you can start learning right away.

Next step is an important step: Writing the proposal

On successfully clearing the previous levels, you with the help of you prospective mentor have to write a proposal for the project. This is the most important step as your application is almost completely judged based on this document.

What is a proposal?

A proposal is a document where you concretely (not vaguely) explain the goal of your project and some information about yourself. This involves:

- The things you want to do with a timeline.
- What problems you may face?
- How do you plan to tackle these?
- Any past open source or relevant experience
- etc.

Let’s see my proposal for example.

GSoC 2022 Project Proposal

Implementing REMD methods in Molly.jl

Submitted by: Jaydev Singh Rao

Email: jaydev19@iiserb.ac.in or jaydevsrao@gmail.com
Website: https://jaydevsr.github.io
Github: JaydevSR, Slack: Same name.

Introduction

Molecular dynamics (MD) is a class of simulation methods that are used to study motion of interacting particles (atom or molecules) with time by solving their equations of motion. Molly.jl is a package in native julia that implements many such molecular dynamics algorithms for study of interacting molecular systems like gases, biomolecules etc. This proposal is about the implementation of Replica Exchange Molecular Dynamics (REMD) method, which is one very important technique of MD simulation, in the Molly.jl package as it lacks this feature at present.

What is REMD?

REMD method is a MD simulation method which involves running several MD simulations of a system at different values of some parameter (eg. temperature, energy etc.) in parallel and exchanging the states from these parallel simulations at periodic intervals (see fig. 1). This leads to the sampling of states which would not be accessible by a single run for a single parameter value. Due to this, REMD is very attractive and in significant use in research due to its ability to escape local minima and explore the configuration space of complex biomolecules to a greater extent compared to classical molecular dynamics simulations.

The implementation of these methods into Molly.jl will provide a very good addition to its overall capabilities. By the end on the proposed project it is expected that temperature based REMD method and some variants of this method will be implemented along with suitable tests and documentation.

The Proposal

The concrete plans that are proposed for this are as follows:

1. Implementation of temperature-REMD method which will run multiple replicas of a MD simulation at **different temperature** values on multiple threads in parallel and exchange

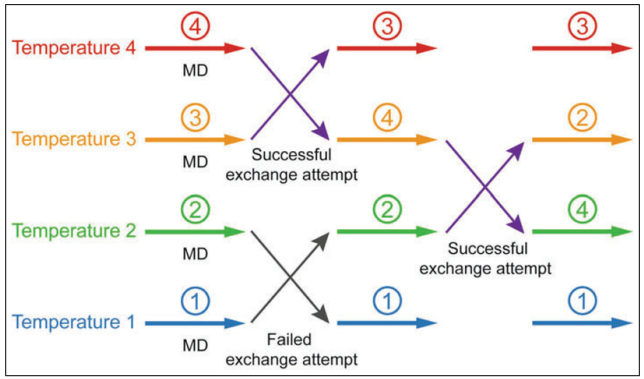


Figure 1: A visual desctription of temperature-REMD simulation.

these at periodic intervals based on Metropolis-Hastings ratios which will give the probability by which such exchange should happen.

Metropolis-Hastings Ratios. *The probability of accepting an exchange between any two states is given by:*

$$P(accept) = \begin{cases} 1 & \Delta < 0 \\ \exp(-\Delta) & \Delta \geq 0 \end{cases}$$

Here, Δ is a quantity that is proportional to energy difference of two states and the difference of inverse temperature.

2. After the implementation is complete, the next step will be to test the implementation on some model systems for which we can get good results from classical MD itself. For example, simulation of Argon using Leonard-Jones potential or simulation of folding of some small protein.

3. One of the integral tasks other that the algorithm implementation itself would be writing sufficient unit tests that cover the complete use cases of temperature-REMD and documenting all the changes and additions made during the complete process.

4. After the successfully completion of the above tasks, a natural next step would be extending the implementation to other REMD methods. Possible variants that can be implemented:

   - Hamiltonian-REMD is the method which is most similar to temperature-REMD and will be easiest to tackle next. It involves running MD simulations of same system at **different energies** instead of temperatures.
   - Replica Exchange Umbrella Sampling (REUS) is an another method which involves adding a harmonic potential term (called umbrella potential) to the original potential energy of the system centered at some different location for each replica. The umbrella potential term prioritizes location where it is centered at, during the MD simulation.

All these methods provide different advantages and can be used in different situations depending upon the biomolecules and the type of process which is being simulated.

Possible hurdles

- The MD simulations in Molly.jl can already utilize multiple threads for better performance. One possible hurdle would be effective use of available threads for MD simulations and simulating multiple replicas which does not sabotage performance as well as accuracy of results.
- Another difficulty can arrive due to the stochastic nature of MD simulations. Due to this the tests should be written in such a way that these are independent of this randomness in results and are not misleading.

Deliverables

Given below is an initial version of the timeline that is to be followed:

- **Phase 1: Before Mid-term Evaluation**
  1. (Before June 13) Understand Molly.jl Codebase.
  2. Implement temperature-REMD.
  3. Test the implementation on model systems (eg. small protein folding, Argon gas).
  4. Add documentation and provide some unit test coverage.
- **Phase 2: After Mid-term Evaluation**
  1. Changes based on feedback from evaluation.
  2. Implementing other variations of REMD.
  3. Test any new implementations on model systems.
  4. Documentation of new changes and maximizing unit test coverage.

What after GSoC?

The experience that I will gain from working on Molly.jl will pave way for future contributions to this package which I am very much interested in and will be grateful for, if I get a chance. I would also like to continue working on anything that is proposed but due to any reason is left uncompleted during the course of the contribution period.

Code Portfolio

There are two projects that I have worked on in the past and am particularly proud of. One is implementing Monte Carlo algorithms i.e. Metropolis-Hastings and Wolff algorithms for studying XY and Ising model (link to project). The other is implementing Numerov-Cooley method, which is a method for solving the Schrödinger's Equation by integrating it and estimating energy by a variational principle (link to project). This is because both of these projects involved learning methods from literature and implementing those from scratch in Julia. This was a really fun and enriching experience for me which is also in line with the requirements of this proposal.

Other than that although I don't have experience contributing to any open source project, as a way of contribution, I have raised issues in the past whenever I felt that those will be constructive to the particular project. Such examples include this and this, both of which have led to PRs in Julia and PlutoUI.jl.

About Me

I am a third year undergraduate student from India. I study Electrical Engineering and Computer Science at IISER Bhopal. I am highly passionate about the field of computational physics in general as it satisfies my intellectual needs from love of both physics and computer science. I have been using Julia since the release of v1.0 and learning it bit by bit it has become **the** programming languge for me. Other than posts on discourse/slack and some forum comments, I have mostly been an external observer in the Julia community. I am motivated towards working on this project because it will give me a chance to take active part in the community that is built around the programming language we all love.

Logistics

During the months of May to July which is summer vacation for me, I have no other commitments except one research project that I have been working on in the past. After that I will have have my semester courses in the remaining period of GSoC (August and September) which I can manage along with this project.

What once you get selected?

First of all, congratulations! You are in for an exciting summer!

The GSoC contribution period consists of three phases:

1. *Community bonding period*: a few weeks of making yourself familiar with the library and the other people in the community.
2. *Coding period (phase 1)*: This is when you start coding and make initial contributions towards your project.
3. *Mid-term Evaluation*: Mentor and contributor give feedback for each other and completion gets you half of your stipend.
4. *Coding period (phase 2)*: You add more to your project and start to finalize it with tests and documentation (very important!).
5. *Final Evaluation*: A week to make some finishing touch-ups and start wrapping up. Then you compile everything you have done and submit it. Successfull completion gets you rest of your stipend.

Stipend rates

There are two project sizes: *medium* and *large*. This is either already given in the ideas list or in case of your own proposal you can decide this with your mentor.

If you are in India during this period, the sipend rates are $1500 for medium sized project and $3000 for large sized project. Other countries have different rates.

Other benefits

Other than the stipend, you get certificates of completion from Google and a special “flair” on your application if you apply to any positions at Google in the future (this was started just this year).

What if you don’t get selected?

This does not necessarily mean you were not good enough! Organisations get funding for fixed number of contributors and due to this they have to make some cuts somewhere.

The next step would be to get feedback on your application from the Org Admin and based on that trying again with a better application next year.

But most important of all!

You don’t need GSoC for making open source contributions. You have to keep contributing if have already started and if not then you have to start contributing with the help of the community. This will go a long way in making your application stronger for the following year.

Now let’s chat!