Phase II Competition - Finalist Application Form

Created: 11/01/2016
Last updated: 11/21/2016

In this form, the six finalist teams for the Open Science Prize are asked to describe the work they have done to develop their prototypes, and make the case for why they should be considered for the phase II Prize.

The information you submit on this form will be considered alongside the prototype you have developed in deciding which team will receive the Phase II Prize.

Please note that, unless otherwise stated, the information you submit on this form will be made available publicly via the Open Science Prize website (under a CC BY 4.0 license), so that it can be assessed as part of the public voting process. All fields, except the final box for additional information are mandatory.

Your application must be completed by 11:59pm Pacific Standard Time on 21 November 2016 . You may edit this form as many times as you like before the deadline.

If you have any questions about the Prize or the review process or if you would like to provide any further information that you would not wish to be made public, please contact David Carr (d.carr@wellcome.ac.uk) or Elizabeth Kittrie (elizabeth.kittrie@nih.gov). Any technical questions regarding this form or the web platform should be directed to (openscience@wellcome.ac.uk).

Page 1

Executive Summary

Please provide a brief Executive Summary of no more than 150 words for the public voting page on the Open Science Prize website. This should be suitable for an informed lay audience, and should briefly describe your prototype and why it should be considered for the phase II Prize. [150 words]

Genome sequences of viral pathogens have the capacity to provide valuable insight into epidemic transmission patterns and viral evolution. But to inform public health interventions in acute public health crises, genomic data has to be analyzed and results disseminated in near real-time. The goal of this project is to promote open sharing of viral genomic data and harness this data to make epidemiologically actionable inferences. For this project, we are developing an integrated framework for real-time molecular epidemiology and evolutionary analysis of emerging epidemics, such as Ebola virus, MERS-CoV and Zika virus. This framework includes an online visualization

platform deployed to the website nextstrain.org that is continually updated as new data becomes available. This platform pools data from across research groups thereby synthesizing disparate datasets and serves to promote open science in the face of public health crises. All source code is publicly available at github.com/nextstrain.

Weblink for prototype

Please provide the public URL for your prototype tool or service (this will be viewed by the public for purposes of public voting):

http://nextstrain.org

Your Prototype

Purpose and need

Please provide a brief summary description of the purpose of the prototype you have developed and the key challenges or needs it is seeking to address [200 words]

Viral pathogens are an enduring threat to global public health and new epidemics are constantly emerging. The ongoing Zika epidemic in the Americas is on course to cause millions of infections and tens of thousands of birth defects. In an emerging outbreak where vaccines are not available, such as the 2013-2016 Ebola epidemic and the current Zika epidemic, the public health response focuses on early diagnosis, contact tracing, isolation, and vector control efforts. This strategy is most effective when transmission patterns and geographic spread are well understood. With the advent of high-throughput sequencing and portable sequencing devices, phylogenetic analysis of viral genomes has become one of the most powerful means of inferring viral epidemiology. To maximize the public health benefit of this approach, open sharing and rapid analysis of viral genome data is essential. To address these challenges, we have developed a prototype system to ingest viral genome sequence data, perform rapid phylogenetic analyses and display results of these analyses on an interactive public website. Nextstrain.org facilitates sharing of data via a standardized database, and incentivizes data sharing through a powerful analysis and visualization platform that carefully attributes data producers.

Please summarise the work you have taken forward to develop your prototype since you were awarded the Phase I Prize in April 2016.

i. Progress

The key milestones achieved and the extent to which the goals and challenges you set out to address in your original application were delivered [400 words]

We have built a working prototype platform to ingest sequence data, perform phylogenetic analyses and upload results of these analyses as an interactive visualization to the public website nextstrain.org. Our platform has three key components.

1. A database (termed fauna) of viral genome data with source code at github.com/nextstrain/fauna. In an outbreak scenario, viral genome data may be uploaded to Genbank or may be shared in pre-publication form via the Virological.org forum or via researcher or government websites. These datasets will have different formats and different metadata encodings. We've written scripts to parse these data and merge datasets into our own fauna database that acts as a single repository for downstream analyses. This database is built with a RethinkDB document store and Python scripts to upload, download and sync data.

2. An application (termed augur) to perform informatic analyses on viral genome sequence data with source code at github.com/nextstrain/augur. Sophisticated statistical and phylogenetic analyses are required to make sense of viral sequence data. Here, we've written a complement of tools to perform sequence alignment, construct time-scaled maximum-likelihood phylogenetic trees, reconstruct mutational events and infer geographic transitions. This is built as a Python application with base classes that can be easily extended for particular viral "builds". This application aims to perform state-of-the-art phylogenetic analyses quickly and flexibly.

3. A web visualization platform (termed auspice) with source code at github.com/nextstrain/auspice. Normally outputs of phylogenetic analyses are static figures. Here, we've written a web application to make a rich interactive visualization of inferences from the augur pipeline. This application shows a time-calibrated phylogeny and allows the phylogeny to be explored to show different metadata, to zoom into clades of interest or to subset the data by geographic region or by study. It displays mutations occurring along the tree and an inference of geographic spread. This application is written in JavaScript and uses React to maintain application state logic and D3.js to aid construction of visual components.

We have written scripts to upload new sequence data to the fauna database, download a merged dataset, run an augur build and push updated analysis results to the website nextstrain.org. We are currently maintaining up-to-date displays of influenza, Ebola and Zika viruses.

ii. Team Contributions

The contributions of the team members to the development of the prototype [200 words]

The original bioinformatic pipeline (augur) and visualization platform (auspice) was developed by core team members Richard Neher and Trevor Bedford during the course of 2015 with specific application to influenza evolution. These pipelines and visualization tools were adapted by Richard and Trevor to work with Ebola virus during summer 2015 and adapted to work with Zika virus in spring 2016. Since selection as finalists for the Open Science Prize, we've engaged in a significant refactor to develop a more flexible and more powerful platform. Richard and Trevor are still leading development, but have brought in student, volunteer and consultant efforts. In this refactor, Charlton Callendar, a student at the University of Washington, built a custom RethinkDB database (fauna) to host viral sequence data. Richard lead the development of a more flexible informatic framework (augur 2.0) with contributions from student Pavel Sagulenko at the Max Planck and volunteer Sarah Murata at the University of Auckland. Student Sidney Bell at the Fred Hutch has extended the codebase to work with dengue virus. The web visualization is undergoing a more significant refactor (auspice 2.0) by consultant Colin Megill to move application state to React and provide a more capable foundation for further development.

iii. Significant Achievements

Any significant achievements or key success metrics you wish to highlight - this might include, for example, numbers of users, key endorsements or engagements with users, new partnerships, external funding, and so forth [200 words]

Our project has already shown remarkable public health utility. Our reports on real-time analyses of influenza virus evolution are now used by the World Health Organization to inform the twice-yearly choice of seasonal influenza vaccine impacting hundreds of millions of vaccinations every year. In summer 2015, in the midst of the West African Ebola epidemic a group led by Ian Goodfellow, Matthew Cotten and Paul Kellam at Cambridge University deployed on-the-ground sequencing to Sierra Leone, and a second group led by Nick Loman at the University of Birmingham deployed embedded sequencing to Guinea. In both cases, groups shared Ebola sequences with us as they were generated and we updated the public website (nextstrain.org/ebola/). These groups were then able to use inferences from the website to identify cross-border transmission events and aid in tracing transmission chains. In 2016, our real-time analysis of Zika evolution (nextstrain.org/zika/) has proved a central source for groups to contribute to without publication priority interfering with data sharing. In particular, efforts to

sequence locally acquired infections in Florida led by Kristian Andersen and efforts to do on-the-ground sequencing in Brazil (ZiBRA project) have shown significant engagement and benefited from rapid phylogenetic feedback.

Learning Points

Please briefly highlight any key learning points you took from the work that you undertook to develop your prototype [200 words max]

In our work on influenza virus, we were able to download well curated sequence data from the GISAID EpiFlu database (gisaid.org). This allowed us to focus on informatic pipelines and visualizations. In making a flexible tool for multiple pathogens, we've realized the importance of working with a variety of data formats and the necessity of data curation. Zika genomes appearing in Genbank often need further metadata curation, which we conduct by reference to the original literature. We've also learned how valuable it is to work directly with data producers rather than working exclusively through public databases. We've found that groups are generally quite happy to share data through the nextstrain.org website. In adapting code to work with a variety of pathogens we've also learned the importance of proper abstraction in software engineering. Writing flexible abstract code is more difficult initially, but allows for faster development down-the-road.

Case for Phase II Prize

Please make the case for why your prototype should be considered for the Phase II Open Science Prize against the following key criteria [100 words each]:

i. Impact

The current and potential future impact of the tool or service in terms of advancing research and generating health and societal benefit

Although the technology now exists to conduct rapid pathogen genome sequencing, the ability to synthesize data across research groups and rapidly convey results to decision makers is lacking. Real-time analysis pipelines and web-based dissemination has the capacity to revolutionize outbreak response and epidemiological investigation. We believe that our project will come to have significant real-world impact (and already has for influenza, Ebola and Zika, see above). Open sharing of source code will also allow the field to progress faster and for methods-focused groups to build on our pipelines, rather than being forced to develop their platforms from scratch.

ii. Innovation

The degree of innovation associated with the tool or service

In our prototype, we are pushing the state-of-the-art forward both in terms of the statistical phylogenetic machinery and also in terms of visualization tools for phylogenetic data. Notably, our analysis pipelines are designed to take minutes-to-hours instead of days-to-weeks as required by commonly used programs such as BEAST. Our investment in interactive phylogenetic visualizations is also novel. Although there are phylogeny drawing libraries such as jsPhyloSVG, these don't allow for extensive interactivity. This allows us to, for example, graphically interrogate a phylogenetic model using ancestral state reconstruction to identify issues of sample contamination or other sources sequencing error.

iii. Utility

The level of demand and utility associated with the proposed service or tool

The nextstrain platform aims to be useful to multiple audiences. One audience is research groups working on viral sequencing and/or phylogenetic analysis. The nextstrain pipeline provides rapid and reusable tools for these groups. Another audience is public health officials, virologists and field epidemiologists. This audience will be less interested in running pipelines themselves and more interested in a detailed look at outputs and visualizations. Finally, there is a lay audience of many thousands interested in epidemics who want a brief overview of the situation, but won't dig into any analyses themselves. Our web-based visualization is ideal for this audience.

iv. Feasibility & Technical Merit

The feasibility and technical merit of the prototype

Common approaches to the issue of creating online phylogenies (used by NCBI and similar databases) is to compute an on-demand phylogeny from a set of sequences chosen by the user. This is necessarily laborious and slow. We've chosen instead to conduct expensive statistical calculations as daily offline builds and serve only derived inferences via a public website. This gives us the "best of both worlds", allowing for advanced computational models with thousands of viral genomes but also allowing near-instantaneous information display. Additionally, we're using state-of-the-art platforms for our database (RethinkDB), bioinformatic analyses (MAFFT, RAxML, TreeTime) and web display (React, D3.js).

Development & sustainability plan

Please briefly describe your vision, and any tangible steps you have taken, to develop your prototype into a sustainable tool or service that advances the goals of open science [400 words]

In our work on Ebola and Zika, we have heard from the European Mobile Lab, USAMRIID and academic research groups that seeing their isolates immediately posted online to nextstrain.org heavily encouraged these groups to share data in a public and timely fashion. We believe that further work on nextstrain.org will continue to foster a 'virtuous cycle' in which aesthetically appealing and scientifically useful analyses encourage rapid and open sharing of pathogen sequence data, which further improves public analyses on nextstrain.org. We will strive to ensure that credit is properly attributed to those generating and sharing genomic data.

Although the technology now exists to conduct rapid pathogen genome sequencing, research groups are still often waiting for publication to publicly deposit sequence data. The reluctance to share data stems from fear of being 'scooped' and having another research group publish an analysis on their data. In working through this with Ebola and Zika, we have observed that a open online tool is less of a publication threat, but still has immediate public health value. Moving away from publication to online dissemination / discussion in the specific context of pathogen outbreaks of public health concern is necessary to harness the revolutionary potential of rapid pathogen genome sequencing.

Correspondingly, all work on nextstrain.org has been completed in an open online fashion. All software components are available through github.com/nextstrain and are released under GNU General Public Licenses. The website nextstrain.org is fully publicly accessible. We are working now to streamline deployment of these software components to allow other groups to build on our pipelines. The entire software pipeline will be instantiated as a Docker container allowing easy deployment by other research groups. Individual analyses will be developed in a modular fashion allowing other groups to pick and choose which analyses best suit their systems or interest. It is our goal to provide a bioinformatic foundation to make evolutionary analyses of emerging pathogens as rapid and frictionless as possible. We hope to foster a community of developers here, where if someone wants to build out a analysis of another virus, it could be incorporated back into the nextstrain.org website and full credit given.

We have received funding from the NIH to continue development of this software platform (R35 GM119774-01 to Trevor Bedford) and plan to continue to invest in its open source development in coming years.

Final comments

Please use the box below to provide any further information you would like to add, that has not been addressed in the questions above [200 words]

N/A