

Poster H-47

Development of an integrated Comparative Genomics environment



Authors:

Hamilton Ganesan (*Bioinformatics, University of Pretoria*)

Fourie Joubert (*Bioinformatics, University of Pretoria*)

Short Abstract: The recent spurt in sequence availability has sparked the era of comparative genomics (CG). Many stand-alone CG tools are available but there is no common environment existing to integrate available software. This project aims to integrate existing CG software into a common environment enabling streamline and efficient CG studies.

Long Abstract:

The output of sequence data and fully sequenced genomes has seen a dramatic rise in the last few years. As a result, researchers are now also looking at comparing genome sequences in order to find similarities and elucidate genome structures. This easy availability of sequence data has been instrumental in sparking the era of comparative genomics (CG). CG is essentially the study of genomes in order to find relationships between them. For instance, our understanding of the human genome has been helped along greatly due to our prior knowledge of the mouse genome and a CG study of the two genomes has shown us how remarkably similar our gene structures are.

Although there is a vast amount of sequence data output, there is not a correspondingly rapid annotation method, herein lies the bottleneck. CG studies have proven to be a useful tool in genome annotation. Although there are a number of CG tools available, there as yet, does not exist a streamlined way of conducting CG studies. Contemporary CG studies are still performed by a manual search, download and installation of appropriate tools for ad hoc purposes. We propose a web-based 'one-stop' comparative genomics environment that allows a user access and usage to a wide selection of CG tools catering for all known needs of a CG study applicable to prokaryotes, eukaryotes and viruses. At present there does exist several integrated CG environments that do to an, extent integrate the various software tools needed for an analysis study e.g homophila, PlasmDB etc, however these systems fall short in several ways. Firstly, most of these 'integrated CG environments' caters for one species in particular. Secondly, the amount of tool usage accessible from them are very limited. Also, these environments don't allow data storage and the ability to interchange data between each of the software tools. This project aims to address all these shortcomings. Typical CG studies require functionality such as whole genome alignment and comparison, domain and motif searching, synteny browsing and gene finding to name a few. A wide range of contemporary open source tools will be integrated into the system to perform these functions. Examples of such tools include Glimmer, GeneMark, MUMmer, MGA and Enterix. The system will provide a user with easy access and usage to these tools via one common user friendly web-interface therefore not requiring any installation on the users part or a great demand on system resources. It must be emphasized that the focus of this project is to not develop and write new analysis software but to maximize the use of existing software by combining them into one environment. The system will also be designed in such a way so as

to make the system easily extensible. Therefore, as new CG needs and respective comparative genomics software tools become available, they would be incorporated into the system with minimal effort and interruption to the other services offered within the module. In keeping with the research focus of our department, as a validation study, we will be looking at plasmodium as the test species, however, the system will be designed to cater for a wide range of species and will make use of the very latest technologies and trends in order to keep abreast of international standards.

This CG module is part of a larger Functional genomics information management system initiative and on completion will provide a host of other functions to individual users and groups such as a structural, sequence, literature mining, microarray and genotyping modules.