ACSESS proceedings
2010

# Real-Time Metagenomics

Daniel A. Cuevas, Joshua A. Hoffman
and Robert A. Edwards

**AP10-11**

CSRC an interdisciplinary gateway to the future
SAN DIEGO STATE UNIVERSITY

# Real-Time Metagenomics

Daniel A. Cuevas[*1], Joshua A. Hoffman[2], Robert A. Edwards[3]
Edwards Bioinformatics Lab, Department of Computer Science,
San Diego State University, 5500 Campanile Dr. San Diego, CA 92183
*Corresponding Author, [1]dcuevas08@gmail.com, [2]keshisaru@gmail.com, [3]redwards@sciences.sdsu.edu

**Abstract:** In the last few years a new technology called metagenomics has revolutionized biology. This technique allows biologists to sequence the DNA (genetic makeup) of all the organisms in an environment. The Real-Time Metagenomics project provides biologists with a variety of tools to annotate metagenomes using web 2.0 technology, including web services (RTMg.web), Google's Android cell phone operating system (RTMg.mob), and all OpenSocial-based social network sites (RTMg.os). This technology allows biologists to perform useful, fast, and easy bioinformatics services on metagenomic data. In fact, annotations of DNA sequences are performed within minutes; something that used to take hours or days.

**Keywords:** metagenomics, android, opensocial

## 1. Introduction

The proliferation of DNA sequences and metagenomic data being collected by biologists has brought many scientists to focus on the fastest and most efficient methods to analyze this data. Recognizing key sequences of nucleotides in a strand of DNA can provide a biologist a great deal of information on the function that given protein has in the organism. This has become an effort for computational and computer scientists to analyze the sequences and provide a simpler method of examining what is in these pools of organisms.

Using our technologies, DNA sequences are annotated quickly in order for biologists to observe and analyze the metagenome. These annotations have created a massive database of proteins, functions, and other properties found in all organisms. The issue of how to process this information and supply it to others in a user-friendly manner now arises.

As a response to the situation, a set of tools was designed in order to accommodate this growing database of metagenomic information. The Real-Time Metagenomics project (RTMg) consists of three core components that provide services to biologists. Some of these services include performing annotations on a FASTA-formatted file of DNA sequences and organizing and displaying the metagenome-annotated data in a user-friendly manner.

Section 2 will briefly explain metagenome annotations. The following three sections (Sections 3-5) will describe each of the RTMg components separately and how they all work seamlessly with one another.

## 2. Annotations

Annotations can be described as assignments of functions and information to the DNA sequence of organisms. This information consists of the analyzed gene functions encoded in the sequences. The functions range from the encoded proteins to the various levels of subsystems those proteins were classified in. This analysis takes into consideration that a set of multiple genomes was used for analysis and annotation, allowing a more accurate determination of which functions are found in the sequences. These annotations are used throughout the RTMg project as they allow users to both retrieve an analysis of their own DNA sequence files and display/examine their annotation results.

## 3. RTMg.web

RTMg.web performs annotations on sequences from any web browser. This application consists of a CGI script which allows a user to upload a FASTA-formatted DNA sequence file [1] to our servers. In return, the annotation results are sent back to the user for examination, sharing, or exportation. The user is able to view the protein functions, along with the different levels of subsystems found in the sequence file. These annotations results can be stored for use throughout the RTMg suite.

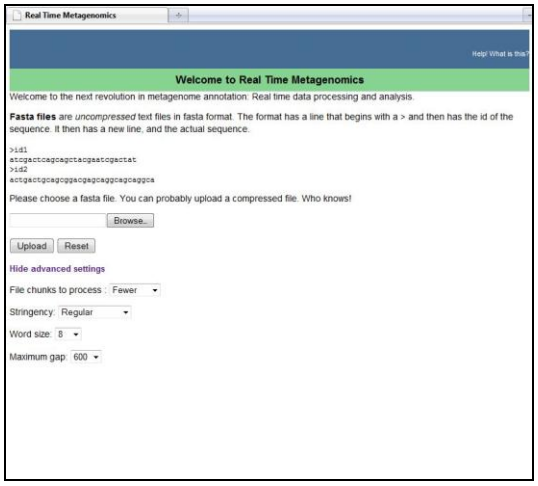



**Figure 1.** Home page of RTMg.web with file upload field and annotation parameters.

### 3.1 RTMg.web Features

There are many features RTMg.web offers, but performing annotations is its primary function. **Figure 1** shows a snapshot of the home page of RTMg.web. An easy to use form is displayed along with other options or parameters the user is able to change. These parameters modify the annotations in different ways: *File Chunks to Process* changes how much the server breaks apart the sequence file to form smaller pieces. This allows the workload of analyzing the sequences to be split up so the annotations can be completed in time; *Stringency* modifies how often the function's nucleotide sequence must appear in the sequence; *Word Size* determines the number of nucleotides that must be present for each function found; *Maximum Gap* determines the number of base pairs allowed between each function's sequence that is found.

After uploading the DNA sequence file, it is processed within just a few minutes. The results are shown to the user in a tabular format shown in **Figure 2.** Depending on what level of subsystem the user is viewing, different information is displayed. The different levels include *Function, One Level of Subsystems, Two Levels of Subsystems, Three Levels of Subsystems,* and *OTU*. After acquiring these results, the user is now able to export or save this data for later use.


**Figure 2.** An example of the annotation results showing *Three Levels of Subsystems*.

## 4. RTMg.mob

With the release of the new Google Android cell phone operating system [2] came the idea to implement this annotation service on these new phones. This Android application allows the user to perform, view, and share their annotations on their cell phone. RTMg.mob concurrently downloads the annotation results while letting the user open other phone applications in tandem. Then after obtaining the annotations, the user is allowed to view the results in different aspects and share them via email or via our server.

This application performs very similar tasks when compared to RTMg.web. It uploads FASTA-formatted files to our server and displays the annotation results back to the user. The obvious difference with this particular application is that it runs on an Android mobile device.

### 4.1 RTMg.mob Features

Because RTMg.mob is very similar to RTMg.web, its key features are also very similar. What separates this from the project is that everything is done on a mobile device. This means that any user can perform any of these functions wherever they are. With the availability of 3G cell phone connection virtually anywhere, this allows someone to quickly annotate their DNA sequence at a moment's notice.


**Figure 3.** The main view of RTMg.mob.

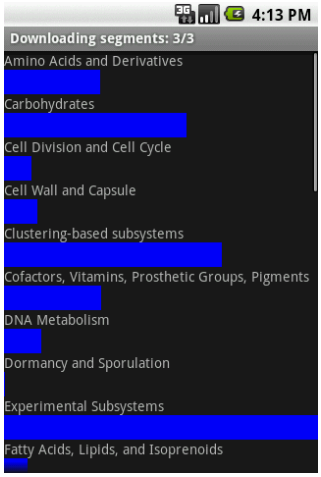



**Figure 4.** A bar graph representing the frequency of all the level one subsystems found in a sample.

**Figure 3** shows the main app view of RTMg.mob. It is consistent with RTMg.web and allows the user to supply the same parameters. Each application is able to cross communicate when retrieving results and performing annotations. Consistency must be kept as everything is using the same format of data.

Because of the limitations with 3G upload/download speed, annotations on a cell phone may take a few minutes longer to retrieve. This brings up an issue regarding waiting for results. We found that it is important that the user be able to leave the application while it concurrently downloads the results. This is another aspect of the Android operating system that RTMg.mob makes use of. With Android's ability to multi-task, the user is able to upload their sequence file for annotation, leave the application to perform some other cell phone function (i.e. browse the web, text message a friend), and then return to the application through a cell phone notification alerting the user the annotations have completed downloading.

Although a notification occurs when the downloading is complete, the user is still able to view their results while the download is still active. **Figure 4** shows an example of one of the different result views available to the user. These results are populated as the annotations are processed and downloaded onto the cell phone. This is important when dealing with very large sequence files spanning a large metagenome. Waiting for the annotations to completely download can be discouraging to most. This app also allows the user to see the development of the results and see how much of the annotations have completed at any given point in time. When the results are finished, the user is able to save the results to the cell phone's SD card for later viewing. There is also the option to email the results or store it within our server for sharing among the rest of the RTMg project.

## 5. RTMg.os

The popularity of online social networks has driven the world to a more connected society where it has become simple and effortless to continuously communicate with a large amount of people. The idea of communicating a single message or thought to hundreds of people in just a few minutes has motivated this area of the project. This OpenSocial-based application [3] focuses on storing and sharing annotated-metagenomic data through our server. It brings a more social aspect to the RTMg project by displaying and sharing every user's data among their network of friends and colleagues.

RTMg.os does not yet perform annotations on DNA sequence files because its main focus is on developing an application with a social aspect for the project. By using an OpenSocial online network, we are given the ability to develop applications for numerous social networks, such as *MySpace* [4] and Google's own *orkut* [5]. Because OpenSocial networks share the same API framework and social functions, applications developed for these networks can be easily and quickly ported to each other. This means that this application, with very few minor

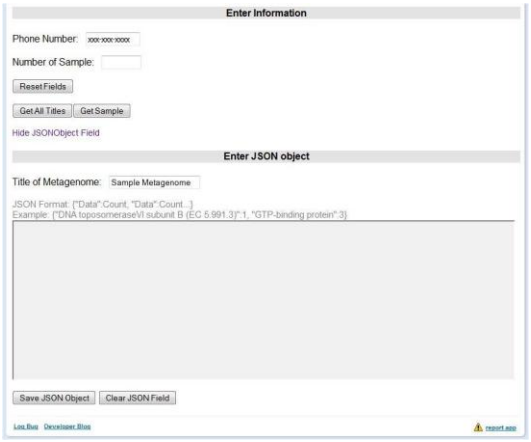
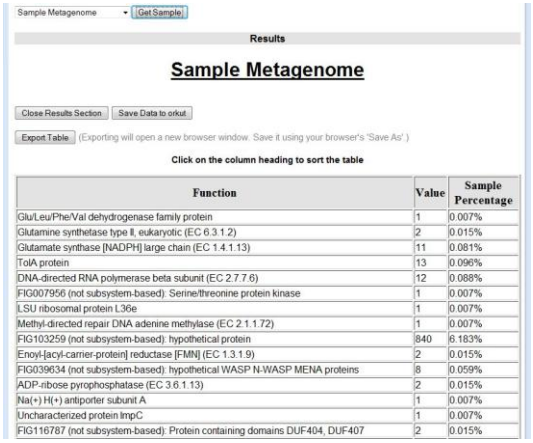



**Figure 5.** Main canvas view with the annotation text field open for saving.




| Function | Value | Sample Percentage |
|---|---|---|
| Glu/Leu/Phe/Val dehydrogenase family protein | 1 | 0.007% |
| Glutamine synthetase type II, eukaryotic (EC 6.3.1.2) | 2 | 0.015% |
| Glutamate synthase [NADPH] large chain (EC 1.4.1.13) | 11 | 0.081% |
| TolA protein | 13 | 0.096% |
| DNA-directed RNA polymerase beta subunit (EC 2.7.7.6) | 12 | 0.088% |
| FIG007956 (not subsystem-based): Serine/threonine protein kinase | 1 | 0.007% |
| LSU ribosomal protein L36e | 1 | 0.007% |
| Methyl-directed repair DNA adenine methylase (EC 2.1.1.72) | 1 | 0.007% |
| FIG103259 (not subsystem-based): hypothetical protein | 840 | 6.183% |
| Enoyl-[acyl-carrier-protein] reductase [FMN] (EC 1.3.1.9) | 2 | 0.015% |
| FIG039634 (not subsystem-based): hypothetical WASP N-WASP MENA proteins | 8 | 0.059% |
| ADP-ribose pyrophosphatase (EC 3.6.1.13) | 2 | 0.015% |
| Na(+) H(+) antiporter subunit A | 1 | 0.007% |
| Uncharacterized protein ImpC | 1 | 0.007% |
| FIG116787 (not subsystem-based): Protein containing domains DUF404, DUF407 | 2 | 0.015% |

**Figure 6.** An example results view after retrieving a previously saved annotation set.

changes to the HTML code, can be installed in each of these OpenSocial-based network containters.

### 5.1 RTMg.os Features

This application is essentially a web application that utilizes the social network environment. **Figure 5** shows part of the main canvas view of RTMg.os. It contains different HTML input fields: *Phone Number* is used as a unique identifier when storing metagenomic data on our server; *Number of Sample* is another identifier to obtain a previously stored metagenomic sample; *Title of Metagenome* is simply the title of the metagenome the user wants to save; the large annotation text field is used to enter in the annotations for storing into our server. These are the four input fields used when attempting to either store or retrieve any metagenomic data into or from our server. The use of a phone number can bring up a few security questions. It is important to know that this number is encrypted before it is stored in our server so the phone number is never saved nor stored in our implementation. The reason we use the phone number is to better synchronize the Android application with the OpenSocial application.

When attempting to store your metagenomic data, a user would enter in their unique phone number, a title for their metagenome, and their annotated results in the JSON [6] format. In this key-value paired format, the function is the key and its count (or number of occurrences in the DNA sequence) is the value. Once the data is saved the application will supply the user with its sample number which is used later for retrieval.

In order to retrieve any saved data, the user has two different options to choose from. They can supply their phone number and click the *Get All Titles* button in Figure 5, or they can enter the phone number and sample number that was given immediately after saving that specific annotation set and then click the *Get Sample* button. Using the first method allows the user to view every sample they have saved on our server through a drop down menu. This method also relieves users of the trouble with remembering sample numbers. Then, by selecting any of the displayed samples the user is able to retrieve their data in a tabular format, shown in **Figure 6**. When the second method is used (using the *Get Sample* button), the results page would display instantly without having to view a drop down menu.

In the results page, the user is able to export and save the data. One very important component of this application is to be able to utilize the social network and the features that come along with it. Thus, another option has been added in the results page that allows the user to share this newly acquired information with their friends or colleagues. By simply clicking the *Save Data to orkut* button that lies right above the table, all of the user's friends that also have RTMg.os installed will be able to see this new metagenomic data.

The main canvas view has an expandable Friends section that will load the all of the user's friends, their *orkut* profile picture, and a drop down of all the annotation samples they saved

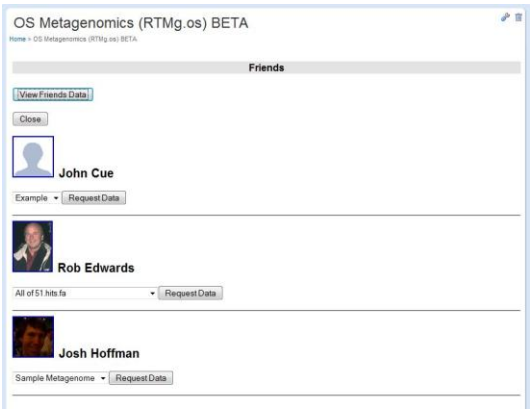



**Figure 7.** The expanded Friends section from the top of the main canvas view.

for sharing, as seen in **Figure 7**. This incorporates the essence of social networking as groups of colleagues and team members can access, store, and share all of their metagenomic data amongst themselves.

## 6. Conclusions

RTMg is a successful model in the Bioinformatic community. Even in its early BETA stages of development, its ability to further connect the scientific community together using a various set of technology can be useful. We have been successful in creating an easy-to-use service by applying it to modes of communication used throughout society today. Smart cell phones are becoming a viable work-related tool for many businesses and now for researchers as well. Social network popularity has spread throughout the world in recent years. With our server holding over 300 publicly available, published metagenomes, access to a massive database of biological information can all be done through the RTMg applications. We have learned that one way to move forward in research is to apply unorthodox and unconventional methods. Implementing research tools in the Android-based cell phone and then in a social network strikes that note very well.

## 7. References

[1] *FASTA format*
http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

[2] *Android Operating System*
http://www.android.com/

[3] *OpenSocial*
http://www.opensocial.org/

[4] *MySpace*
http://www.myspace.com/

[5] *orkut*
http://www.orkut.com/

[6] *JSON format*
http://www.json.org/