# Bioinformatic Resources for Genomic Information Retrieval on Distributed Computing Approach

**Wagner Arbex[1], Leonardo Carvalho Nápolis Costa[1,2], Leonardo Mariano Gravina Fonseca[1], Camillo de Léllis Falcão da Silva[2]**

[1]Brazilian Agricultural Research Corporation – Embrapa
Juiz de Fora, MG, Brazil

[2]Federal University of Juiz de Fora – UFJF
Juiz de Fora, MG, Brazil

wagner.arbex@embrapa.br

***Abstract.*** *The National Center for Biotechnology Information (NCBI) website provides computational resources from several web services, which can be reached with some programming languages and/or using specific modules of theses programming languages. Nevertheless, almost all services must be accessed from knowledge of metainformation about the data. Eventually, without knowing any information about what one wants to search, such as a single DNA sequence representing "string" by computational view, hindering access to NCBI's resources. This paper presents a bioinformatics resources to enable metainformation discovery of raw genomic data such as genetic sequences without any knowledge about them.*

## 1. Introduction

The NCBI keeps the largest data repositories and bioinformatics computing resources worldwide, and creates and maintains automated systems for storing and analyzing knowledge about molecular biology, biochemistry and genetics, thus facilitating the use of databases and software for searching the scientific community [NCBI 2013].

The issue featured occurs when it is necessary perform searches without knowing any information about the data to be studied, because almost all NCBI's search services are based on some knowledge about the investigated data. This paper presents a solution for cases where the data investigated is raw nucleotide sequences and which are not known any information, even the organism of origin of sequences.

The Entrez system [NCBI 2011] is an example of NCBI's resources, which is a search and retrieval system of NCBI that offers users integrated access to many features and data structures..

From of the NCBI unique identifier[1] of data, can be reached innumerable information about the them. All information about a DNA sequence can be retrieve using Entrez resources, as in the Fig. 1[2], it can be seen which was recovered from

[1]The NCBI unique identifier, e. g., is the "accession number", which is "given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ)" [NCBI 2002].

[2]The sequence of characters (...) indicates information edition to readability of the figure.

the query [Sayers 2011]:

`http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=34577062&rettype=fasta&retmode=text`

In the query were informed identification parameters on what you want to look for, and the part **id=34577062** determines a **GI** number[3]. In the specific case, only in this way it was possible to retrieve other information about the sequence, including the sequence itself.

```
>gi|34577062|ref|NM_001126.2| Homo sapiens (...), mRNA
GGAAGGGGCGTGGCCTCGGTCCGGGGTGGCGGCCGTTGCCGCCACCAGGGCCTCTTCCTGCGGGCGGTGCTGCCGAGGCCGGCCTGCGCGGGGC
AGTCATGGTACCCCCTTGAGCGGGCTGTGGCGGAGAGCGGGGCGGGGACTGGCTGGAGGGTGGCGGCCCGGCGGGGCGGGGGCGGGGCCGGCCT
CTGGCTCCTTCTTCCTCTGCATGTGGCTGGCGGCCGCAGAGCAGTTCAGTTCGCTCACTCCTCGCCGGC(...)
```

**Figure 1. Retrieved information from Entrez query perform.**

## 2. Materials and Methods

If there was no knowledge of these sequences, the BLAST (Basic Local Alignment Search Tool) software could be used to find the first information about them. The BLAST software is able to sequence alignment and find similarity between them. Seting up similarity between sequences is a powerful tool for identifying the unknowns in the sequence world [Korf et al. 2003].

There are three ways to use the BLAST software: directly through NCBI's website, in the BLAST page, when it is possible submit a sequence or a list of sequence, "putting" one by one, or read all sequences from a FASTA file; the second way, it is necessary to get a version of BLAST program, e. g., available from NCBI website, and the particular BLAST file to the aligning process. To install the BLAST program on a local machine to run it using the BLAST file. In the third one, the BLAST process is running by remote call, like a remote procedure call (RPC) mode.

Perl is a programming language originally developed to text handling, and today beyond keep all of features it is used for numerous kinds of different applications [Christiansen et al. 2012, Perl.org 2013b]. The CPAN (Comprehensive Perl Archive Network) website repository, and its website stores thousands Perl modules, for several distribuitions [Perl.org 2013a] to be used by any Perl developer. In the CPAN repository is possible to find several resources to build solutions for accessing remote database or others computational resources, including modules for distributed computing.

The best resource provided from Perl to bioinformatics and computational biology is the BioPerl project [Stajich et al. 2002, BioPerl Core Developer 2012] which is an "collaboration of biologists, bioinformaticians, and computer scientists (...)" [Stajich et al. 2002] to developing of a "toolkit of Perl modules". The BioPerl has several procedures to able access NCBI's services, and it is can run the BLAST program from remote call. Therefore, from Perl it is possible build a distributed system in the architecture client/server, executing remote calls, over HTTP protocol.

[3]The **GI** number is "Genbank identifier", which is a kind of NCBI unique identifier.

To run the task to get metainformation about sequences, from NCBI website, without any a priori information on it, was developed blastRemoteCall.pl Perl script, and the adopted approach strategy was to get unknown sequences in a the FASTA format and submitted them to NCBI's BLAST program from remote call.

To code blastRemoteCall.pl script was used Perl language with BioPerl modules to call remote BLAST. In particular, in the blastRemoteCall.pl source code, were used `Data::Dumper` module and the BioPerl modules `Bio::SearchIO` and `Bio::Tools::Run::RemoteBlast`. The `Bio::Tools::Run::RemoteBlast` module is the most important for script strategic role, because the method which performs the remote BLAST is implemented in it.

To run the remote BLAST is needed set parameters. These parameters assign, among others features, the "kind" of BLAST and which database will be used. In the blastRemoteCall.pl script, these parameters received **blastn** and **nr** values, specifying the BLAST program and database for nucleotides, respectively. Next, the BLAST remote submission is done with each sequence to be investigated sent in FASTA[4] files (Fig. 2) as a submission parameter.

```
>
GGAAGGGGCGTGGCCTCGGTCCGGGGTGGCGGCCGTTGCCGCCACCAGGGCCTCTTCCTGCGGGCGGTGCTGCCGAGGCCGGCCTGCGCGGGGC
AGTCATGGTACCCCCTTGAGCGGGCTGTGGCGGAGAGCGGGGCGGGGACTGGCTGGAGGGTGGCGGCCCGGCGGGGCGGGGGCGGGGCCGGCCT
CTGGCTCCTTCTTCCTCTGCATGTGGCTGGCGGCCGCAGAGCAGTTCAGTTCGCTCACTCCTCGCCGGC(...)
```

**Figure 2. FASTA sequence file sample for BLAST remote submission on blastRemoteCall.pl script**

## 3. Results and Approach Analysis

The paper proposes remote access on NCBI's website to use their web services and perform BLAST program to get information about unknown sequences in FASTA format. Particularly, these sequences could be nucleotide sequences, as a DNA or RNA sequences.

The means adopted to implements this strategy was process remote calls as a RPC. To implement the RPC was used modules of BioPerl project to set up the BLAST parameters; executes and submits the BLAST program to mining the database; and retrieve the answer.

The send and receive procedures interaction – *message passing* between them – defines the client/server communication. The client and server doing *synchronous communication* – client waits for the server response – or they doing *asynchronous communication* – the client performs a request and continues processing, without waits for the reply [Tanenbaum and Van Steen 2006]. The blastRemoteCall.pl script implements asynchronous communication mode.

Therefore, as the BLAST server is able to receiver and handle each request independently of the others and is not necessary for it waits for the reply, then the entire process becomes more efficient. The implemented communication method

[4]The definition line – marked with '>' – is empty, because there is no information about the sequence.

with distributed system concepts creates a lightweight and efficient message passing between the Perl script, client application, and the server on NCBI's website.

## 4. Conclusions and Perspectives

Search for metadata about sequence fragments can be necessary many times on bioinformatic and computational biology tasks and the NCBI website is an important place to do prospecting like these, providing efficient tools. However, almost all NCBI's resources need of a identifier for the sequence which look up its metadata. Nevertheless, it is not uncommon encounter sequences or fragments of sequences without identifiers, because, many times only knows the organism origin sequences and no more information or knowledge about them.

This paper introduces and analyzes the first version of an intelligent retrieval for metadata of unknown sequences with a Perl script using Bioperl and other modules as well as resources of the NCBI website. The solution proposed shows another approach and the search for metadata is done from remote execution of the BLAST alignment program, enabling the recovery all the information found by search of similarity among sequences.

## Acknowledgments

The authors thanks to reviewers who gave useful comments, and would like to express thanks to the State of Minas Gerais Research Support Agency (FAPEMIG) for the partial support for the accomplishment of this paper.

## References

[BioPerl Core Developer 2012] BioPerl Core Developer (2012). BioPerl. Website.

[Christiansen et al. 2012] Christiansen, T., Foy, B., and Wall, L. (2012). *Programming Perl*. O'Reilly Media, Inc., Sebastopol, 4 edition.

[Korf et al. 2003] Korf, I., Yandell, M., and Bedell, J. (2003). *BLAST*. O'Reilly & Associates, Inc., Sebastopol.

[NCBI 2002] NCBI (2002). *The NCBI Hanbook: Glossary*. NCBI, Bethesda.

[NCBI 2011] NCBI (2011). Entrez help: NCBI bookshelf.

[NCBI 2013] NCBI (2013). Our mission.

[Perl.org 2013a] Perl.org (2013a). The comprehensive Perl archive network. Website.

[Perl.org 2013b] Perl.org (2013b). The perl programming language. Website.

[Sayers 2011] Sayers, E. (2011). E-utilities quick start. In McEntyr, J. and Ostell, J., editors, *Entrez Programming Utilities Help*. NCBI, Bethesda.

[Stajich et al. 2002] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618.

[Tanenbaum and Van Steen 2006] Tanenbaum, A. S. and Van Steen, M. (2006). *Distributed systems: principles and paradigms*. Prentice Hall, 2 edition.